EVALUATING EXPERIENTIAL LEADER DEVELOPMENT: A PROGRAMMATIC EVALUATION AND COMPARISON OF THE EFFECTIVENESS OF US AIR FORCE SQUADRON OFFICER SCHOOL CURRICULA

THESIS

Jeffrey G. Holland, Captain, USAF

AFIT/GLM/ENV/08-M01

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY
AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

The views expressed in this thesis are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the United States Government.

AFIT/GLM/ENV/08-M01

EVALUATING EXPERIENTIAL LEADER DEVELOPMENT: A PROGRAMMATIC EVALUATION AND COMPARISON OF THE EFFECTIVENESS OF US AIR FORCE SQUADRON OFFICER SCHOOL CURRICULA

THESIS

Presented to the Faculty

Department of Operational Sciences

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the Requirements for the

Degree of Master of Science in Logistics Management

Jeffrey G. Holland, BS

Captain, USAF

March 2008

AFIT/GLM/ENV/08-M01

EVALUATING EXPERIENTIAL LEADER DEVELOPMENT: A PROGRAMMATIC EVALUATION AND COMPARISON OF THE EFFECTIVENESS OF US AIR FORCE SQUADRON OFFICER SCHOOL CURRICULA

Jeffrey G. Holland, BS
Captain, USAF

Approved:

_____--signed--_____	13 Mar 08
Kent C. Halverson, Lt Col, USAF (Chairman)	date
_____--signed--_____	13 Mar 08
Pamela S. Donovan, Lt Col, USAF (Member)	date
_____--signed--_____	13 Mar 08
Alexander J. Barelka, Lt Col, USAF (Member)	date

Abstract

Leader development programs often employ experiential learning exercises. The impact of such exercises is not clear. This research investigated experiential leader development using a quasi-experimental design to analyze the differences in two consecutive US Air Force Squadron Officer School (SOS) in-residence classes. The curriculum was altered between classes by the addition of the Combat Leadership Exercise (CLX), an experiential war-gaming activity.

Experiential programs regularly use mean differences between pretest and posttest measurements to represent program impact. However, research shows that participants may change the way they evaluate themselves between test administrations due to their experiences in the programs, a phenomenon known as response shift. Response shift renders results of mean differences evaluation invalid.

The common means differences showed SOS had weak impact on leader development and showed no difference between the treatment class (CLX) and the comparison class (no CLX). However, structural equation modeling identified the presence of response shift within each SOS class, indicating that students had reconceptualized or recalibrated certain aspects of leadership measured before and after SOS.

The implications of response shift and its measurement are discussed. An argument for changing the leader development evaluation paradigm to legitimize response shift as a program outcome is presented.

"Leadership and learning are indispensible to each other."
– President John F. Kennedy, 22 Nov 1963

To my wife & son

Acknowledgments

I would like to express my sincere appreciation to my faculty advisor, Lt Col Kent Halverson, and Lt Col Alexander Barelka for their guidance and support throughout the course of this thesis effort. Their insight made this research possible.

I also would like to thank Dr. Arden Gale and Maj Kenneth Shugart, my liaisons and champions at Squadron Officer School, who enabled my survey administrations and were always willing to assist me. Most of all, I offer my appreciation to the members of SOS classes 07G and 08A who allowed me to impose on their SOS experience.

Jeffrey G. Holland

Table of Contents

List of Figures

List of Tables

EVALUATING EXPERIENTIAL LEADER DEVELOPMENT: A PROGRAMMATIC EVALUATION AND COMPARISON OF THE EFFECTIVENESS OF US AIR FORCE SQUADRON OFFICER SCHOOL CURRICULA

I. Introduction

Interest in leadership has never been greater. Stories of successes and failures in leadership populate news outlets daily. Additionally, organizations increasingly see leadership as a source of competitive advantage (Conger & Benjamin, 1999; Day, 2001; Kouzes & Posner, 2002; McCall, 1998; Phillips & Schwartz, 2004; Vicere & Fulmer, 1998). With leadership prominently featured in media and its presence coveted by organizations, it follows that leadership development is the focus of much attention (Conger & Benjamin, 1999; Day, 2001; Kouzes & Posner, 2002; Phillips & Schwartz, 2004). However, "leadership theorists and practitioners often disagree about what leadership is, how leaders behave, what makes a good leader, and what effective leader performance looks like" (Martineau, 2004: 234). This discord makes it difficult to create and evaluate leader development programs. Thus, those charged with putting leader development into practice are faced with three daunting questions: what aspects of leadership will the program focus on, how will the program be delivered, and how will the program be evaluated?

The rift between the academic and practitioner communities rapidly reveals itself in the most basic investigation of leadership. A Google™ search of "leadership" returned

over 152 million hits, over 61 thousand of which were books; when the search is limited to Google™ Scholar, the same search returned over 2.8 million hits (1 Feb 2008). Approaches to leadership range from the traditional individual-focused theories to more integrative intra- and interpersonal interactions to wholly interpersonal perspectives (Bass & Stodgill, 1990).

The variety of opinion increases when the issue is broadened to consider not only leadership but leader development. A Google™ search of "leader development" returned over 22 million hits, over 15 thousand of which were books; limited to Google™ Scholar, the same search returned nearly 750 thousand hits (23 January 2008). While there is surely significant overlap among the hits, the search results make it clear that several opinions on leader development exist and the number of results returned across the entire web versus those found only in the scholarly realm indicate the topic is more common among practitioners than academics.

Adding to the confusion for those charged with creating leader development programs is the need to choose from multiple potential delivery methods. The choice of delivery method is crucial to program success because the delivery method can influence the transfer of the desired insight, skills, and attributes in leader development (Conger & Benjamin, 1999; McCauley & Van Velsor, 2004; Vicere & Fulmer, 1998). A widely popular delivery method is experiential learning, sometimes referred to as outdoor or adventure learning (Albertson, 1995; Buller, Cragun, & McEvoy, 1991; Hattie, Marsh, Neill & Richards, 1997; Hernez-Broome & Hughes, 2004; Judge, 2005; Roland, 1984; Ronan, 2003; Useem, Davidson, & Wittinberg, 2005; Wagner, Baldwin, & Roland,

1991). However, applications of experiential learning conducted in environments outside of the traditionally envisioned austere or wilderness settings are evident.

Administrators and advocates of experiential learning programs strongly assert program effectiveness in promoting personal development (Baldwin, Persing, & Magnuson, 2004; Useem et al., 2005). According to Kolb's (1984) experiential learning theory (ELT), the dualistic approach to knowledge acquisition and transformation taken by these programs significantly heightens the effectiveness of the programs and better promotes personal development (Kayes, 2002; Useem et al., 2005). Yet, empirical support for these claims is lacking (Buller et al., 1991; Garvin, Nason, & Otto, 1996; Keller & Olson, 1990; Roland, 1984; Sheard & Golby, 2006; Wagner et al., 1991; Useem et al., 2005). The dearth of evidence supporting ELT is especially apparent in the field leader development (Useem et al., 2005). This research will begin to fill this void through the investigation of the effect of experiential learning in leader development.

Specifically, this research will examine the effects of experiential learning in a military leader development context. The US Air Force Squadron Officer School (SOS), the second tier of the Air Force's two-school tactical leader development program, recently adopted a new experiential learning activity in an effort to better fulfill its mission "to develop dynamic Airmen ready to lead Air, Space, and Cyberspace power in an expeditionary warfighting environment" (http://sos.maxwell.af.mil). A new activity, the Combat Leadership Exercise (CLX), which is a simulated combat experience, was added to the curriculum in the first class of Fiscal Year 2008. In accordance with ELT, the insertion of an experiential activity into the curriculum should increase the learning

experienced in the program and result in an increase in attendee's leader development (Burke & Day, 1986; Collins & Holton, 2004; Kolb, 1984; Sullivan & Kolb, 1995).

By measuring the leader development experienced in SOS, it is possible to evaluate the effectiveness of SOS as a leader development program. Comparison of program effectiveness before and after the addition of the CLX will quantitatively define the contribution of the CLX to the program's effectiveness, thereby providing insight into the development attributable to the experiential learning exercise.

II. Literature Review

Leader Development

The traditional conceptualization of leadership is as a collection of individual-level traits, attributes, and skills (Conger & Benjamin, 1999; Day, 2001; Day & Halpin, 2004). In this conceptualization, training is naturally limited to the development of intrapersonal attributes and the individual acquisition of skills and abilities (Day, 2001; McCauley & Van Velsor, 2004). McCauley & Van Velsor (2004), expand on the traditional view and define *leader development* "as the expansion of a person's capacity to be effective in leadership roles and processes" (2). Application of this definition moves the developmental focus beyond the individual and includes the interpersonal skills and attributes necessary to facilitate setting direction, creating alignment, and maintaining commitment in groups of people who share common work" (McCauley & Van Velsor, 2004: 2). Thus, *leader development* can be conceptualized to include the development of both human and social capital within an individual and within an organization.

The inclusion of both human capital and social capital in leader development is important. Researchers have noted the limitations of adhering to a purely individual focus (Brass & Krackhardt, 1999; Day, 2001; Fiedler, 1996; McCauley & Van Velsor, 2004). The composition, structure, and function of modern organizations do not allow an individual to "accomplish leadership tasks by virtue of their authority or their own leadership capacity" (McCauley & Van Velsor, 2004: 21). Instead, individuals must build on their own competencies with interpersonal competencies, such as the ability to

generate commitment, inspire trust, and garner respect (Day, 2001). The inclusion of both intra- and interpersonal competencies in leader development requires the developmental process to be both differential and integrative (Day, 2001), with differential referring to a focus on intrapersonal development and integrative referring to the exploitation of interpersonal relationships to achieve results.

The differential aspect of leader development exists in a program to promote the individual's acquisition of an enhanced self-understanding and improved personal skill-set (McCauley & Van Velsor, 2004). The development of such individual-level skills and attributes are a necessary prerequisite in preparing an individual to capture the social capital required for organizational leadership (Day & Halpin, 2004). Built upon the traditional conceptualization of leadership, the preponderance of organizational leadership research focused on this foundation of individual knowledge, skills, and attributes, or KSAs (Day, 2001; Conger & Benjamin, 1999).

Yet, an individualized focus fails to capture an important aspect of leader development. Integrative development requires an individual to build upon individual KSAs and learn to form and exploit interpersonal relationships as a means of achieving a desired end-state or outcome. The value of interpersonal relationships as a component of leadership is well documented (Bass & Stodgill, 1990; Conger & Benjamin, 1999; Fiedler, 1996; Day & Halpin, 2004). The contribution of relationships in learning leadership is also well-established (Kram, 1985; Kram & Isabella, 1985; McCall, Lombardo, & Morrison, 1988; McCauley & Hughes-James, 1994). The known importance of relationships in both the learning and application of leadership makes the

development of interpersonal relationships and the acquisition of interpersonal KSAs a necessary component in an effective leader development program.

Leader Competency Models

In an attempt to capture the differential and integrative aspects of leader development, several organizations rely on competency-based programs in which the competencies cover both intra- and interpersonal skills. Though some may argue against a focus on competencies in leader development (e.g. Mintzberg, 2004; Conger, 2004; Raelin, 2004), Lombardo & Eichinger (2002) concluded that 85% of the competencies needed for effective management are the same for all jobs. With the commonality of competencies required for success, it is understandable that, in the quest to develop effective leaders, many organizations employ competency models as standards of leadership (APQC, 2000; Conger & Benjamin, 1999; Hernez-Broome & Hughes, 2004).

The Air Force Leadership Model

The US Air Force is among those organizations that adopted a competency-based leadership model. The Air Force leverages the experience of its senior leaders and identifies 16 "enduring leadership competencies," or KSAs, in Air Force Doctrine Document (AFDD 1-1), Leadership and Force Development (2006). The KSAs are broken into three key areas: personal, people/team, and institutional (Table 1). These 16 KSAs are those that the Air Force deems essential to effective leadership, and therefore, "should be common to all Air Force members" (US Air Force, 2006: 10).

The Air Force used these enduring leadership competencies as the foundation for a leader development model, known as the Air Force Leadership Development Model (http://www.clrexec.com/site.cfm?id=91 or https://www.afsl.hq.af.mil/fd/fdld). Though each of the KSAs identified in AFDD 1-1 are necessary to be an ideal leader, the degree of use each KSA receives varies as leaders progress upward in the organization (2006). In the Air Force Leadership Development Model, the mix of competencies required is grouped into three levels: tactical, operational, and strategic (https://www.afsl.hq.af.mil/fd/fdld).

At the tactical level, that occupied by first-line supervisors and managers, the Air Force emphasizes the use of KSAs found in the personal area, those that are conducted internal to the leader or "face-to-face" with others. As a leader's scope of responsibility broadens through promotion to operational level, the KSAs in the people/teams grouping increase in importance while those in the institutional group begin to develop in earnest. Finally, while still utilizing the KSAs developed in tactical and operational service, those leaders at the strategic level focus on the KSAs found in the institutional group so they may effectively lead the organization (US Air Force, 2006).

Table 1.

US Air Force Enduring Leadership Competencies

Personal Leadership	***Leading People/Teams***	***Leading the Institution***
Exercise Sound Judgment	Drive Performance through Shared Vision, Values, and Accountability	Shape Air Force Strategy and Direction
Adapt and Perform Under Pressure	Influence through Win/Win Solutions	Command Organizational and Mission Success through Enterprise Integration and Resource Stewardship
Inspire Trust	Mentor and Coach for Growth and Success	Embrace Change and Transformation
Lead Courageously	Promote Collaboration and Teamwork	Drive Execution
Assess Self	Partner to Maximize Results	Attract, Retain, and Develop Talent
Foster Effective Communication		

Note: Adapted from AFDD 1-1, page 11.

Experiential Learning

The origins of the field of experiential learning are generally traced to Kurt Hahn's Gordonstoun School in northern Scotland (Hattie et al., 1997; Weigand, 1995). The school, whose motto is "Plus est en Vous" or "More is in You," was "dedicated to the development of a student's inner resources versus physically and mentally demanding outdoor experiences" (Weigand, 1995: 2). Hahn's model gained traction in the United States during the 1930s through the Civilian Conservation Corps (Judge, 2005). During World War II, the school temporarily relocated to Wales and established a corollary

program to support the training of British sailors. The principles of Gordonstoun and the curriculum of the seamanship and survival class served as the basis of the organization known today as "Outward Bound" (Hattie et al., 1997: Weigand, 1995).

In the aftermath of World War II, Hahn's model spread further in the US through the US military's professional military education schools (Weigand, 1995). By the early 1960s, the Colorado Outward Bound School formed and specialized experiential learning activities emerged in US military academies, such as the US Air Force Academy's Group Reaction Course (Weigand, 1995; Garvin et al., 1996). Experiential learning organizations and programs continued to proliferate during the 1960's and 1970's, resulting in multiple presentations on the subject at both the American Society for Training and Development and the Association for Experiential Education conferences in the early 1980s (Weigand, 1995). Interest in the experiential learning field exploded in the mid-1980s and its effectiveness as a corporate training tool has been in debate since (Weigand, 1995).

Experiential Learning Theory

In 1975, David Kolb attempted to explain the learning process of experiential learning through the Experiential Learning [Theory] (Judge, 2005). Kolb's model, based on "Dewey's philosophical pragmatism, Lewin's social psychology, and Piaget's cognitive-developmental genetic epistemology" (Kolb & Boyatzis, 2000: 2), posits that learning involves the interplay between two independent dimensions of knowledge: acquisition and transformation (Kayes, 2002).

In accordance with ELT, knowledge acquisition occurs through two related, though opposed, means: apprehension (concrete experience) and comprehension (abstract conceptualization) (Kolb, 1984; Sullivan & Kolb, 1995; Baker, Jensen, & Kolb, 2002; Kayes, 2002). Apprehension requires the acceptance of new knowledge through a direct experience while comprehension occurs through understanding of abstract concepts (Kayes, 2002).

Similarly, ELT identifies two methods of knowledge transformation: intention (reflective observation) and extension (active experimentation) (Kolb, 1984; Sullivan & Kolb, 1995; Baker, Jensen, & Kolb, 2002; Kayes, 2002). Intention involves the internal processing of experience while extension requires interaction with the environment (Kayes, 2002).

In identifying these dimensions of acquisition and transfer, ELT claims to identify the learning cycle, or "the whole process whereby knowledge is created through the transformation of experience" (Kolb, 1984: 41), shown in Figure 1. The process begins with a concrete experience. The lessons of the concrete experience are processed through reflective observation. Through this reflection, abstract concepts resulting from the experience emerge and serve as knowledge to be input into future actions. These concepts are then tested through active experimentation and the learning cycle is re-entered. (Kolb, 1984; Sullivan & Kolb, 1995; Baker, Jensen, & Kolb, 2002)

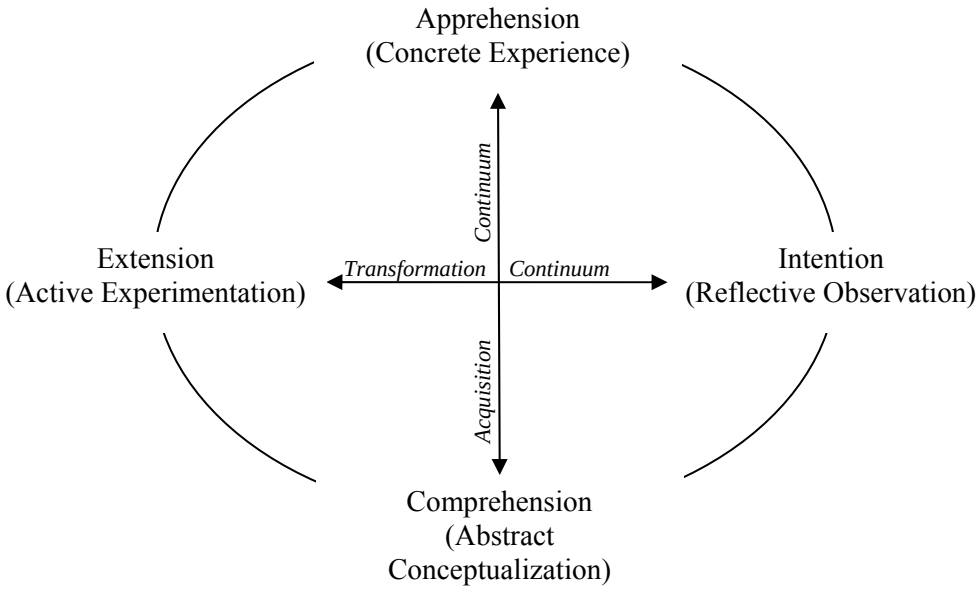


Figure 1. Kolb's Experiential Learning Theory (ELT)

The fundamental difference between classroom and experiential learning is ELT's dualistic approach to knowledge acquisition and transformation (Kayes, 2002). Traditional classroom efforts focus on the comprehension side of the knowledge acquisition continuum identified in ELT through the abstract conceptualization of symbols, such as mathematical operators or written text (Garvin et al., 1996). While this greatly increases the speed at which the material can be presented, Baldwin & Ford (1988) suggest that "it is only effective if students can learn and apply general principles from the lessons to generate action across many similar but unique situations" (as cited in Garvin et al., 1996: 2). The ability to measure such effectiveness is hampered because evaluation of traditional instruction also relies on symbolic learning, as the transfer of knowledge is measured through written tests without the opportunity for application and experience (Garvin et al., 1996). This contrasts with ELT in which action is encouraged

in order to develop generalized principles to apply across diverse situations (Garvin et al., 1996; Kayes, 2002; Judge, 2005).

While ELT is not the only learning theory to propose a learning cycle based on dual continuums, it is among the only such theories that are "both comprehensive and fully generalized" (Kayes, 2002: 140). Given ELT's claims of generalizability and its foundation in the belief that everyone can grow and learn (Miettinen, 1998), its appeal is understandable. Indeed, ELT forms the basis of most experiential learning programs' claimed effectiveness (Kolb, 1999; Hattie et al., 1997; Hernez-Broome & Hughes, 2004; Judge, 2005; Useem et al., 2005).

Experiential Leader Development

The action context of experiential learning makes it an appealing basis for leader development programs. By its nature, experiential learning allows for programs to deliver on each of the necessary factors for leader development programs proposed by Gardner (1990). Experiential learning provides opportunities "to experience the shared responsibilities of group action," tests of personal judgment, exposure to different perspectives, and places participants in "the untidy world, where decisions must be made on inadequate information" (Gardner, 1990: 168)

While the "untidy world" mentioned by Gardner (1990) refers to the chaos of the reality contrasted to the controlled predictability of the classroom, it surely includes the wilderness-based settings common in Outward Bound programs. Indeed, public perception of the adventure-based Outward Bound programs centers on the programs' capability for leader development (Richards, 1975 as cited in Hattie et al., 1997). Stolz

(1992) found that the public's perception was justified; promotional material for Outward Bound programs explicitly claimed to improve teamwork, communication and leadership skills.

Sharing a foundation in Hahn's model with Outward Bound (Garvin et al, 1996), the US Air Force Academy's Leadership Reaction Course (LRC) focuses on strengthening the "five skills required for successful officership in the United States Air Force: leadership, followership, teamwork, communication, and problem solving" (Garvin et al., 1996: 15). The LRC achieves these results through an experiential approach consisting of a pre-brief, 12 situational tasks followed by immediate reflection, and an end-of-course reflective review (Garvin et al., 1996).

The LRC serves several populations: cadets involved in Basic Cadet Training, students in an upper-level academic course (Behavioral Science 310: Leadership Concepts and Application) and non-cadet groups, such as the US Air Force Space Command Lieutenants' Leadership Program and the inner-city youth Reach for Tomorrow program (Garvin et al., 1996). For each population, the purpose of the program remains constant: "successful completion of the task requires teamwork and mental and physical exertion, and may push the group and/or individuals beyond their previously known limits. In addition, the course also provides [participants with] motivation, [a] sense of accomplishment, and positive reinforcement of effective teamwork" (Garvin et al., 1996: 82).

The internal reward attributed to overcoming the challenges of the LRC is not unique among experiential programs. The Wharton Center for Leadership and Change Management at the University of Pennsylvania sponsors the Wharton Leadership

Ventures program, a series of out-of-class exercises that allow students to "directly witness and experience leadership decision making" (Useem et al., 2005: 162). In order to do this, Wharton Leadership Ventures takes students into unfamiliar settings, such as the Antarctic and Patagonia, and attempts to tie lessons available in these environments to concepts and principles conveyed in the classroom (Useem et al., 2005).

In these extreme settings, Wharton Leadership Ventures seeks to build students abilities in four key areas required for effective decision making: transcending self-interest, determining direction, living by one's principles, and acting decisively (Useem et al., 2005). According to Useem et al. (2005), the unique experiential approach of the program overcomes the insufficient delivery of the classroom and allows Wharton Leadership Ventures to be "instructively memorable" and "analytically informed," thereby producing students that better grasp and retain leadership lessons than those students limited to classroom learning (162).

This perspective is shared by Judge (2005) in a review of an experiential leader development course for an eMBA program at the University of Tennessee. Though much of the article reviews shortcomings in previously attempted experiential exercises, which included a mountain trek and a challenge course, Judge (2005) declares that, even when only partially successful, the experiential exercises are superior to traditional leadership programs. Indeed, Judge (2005) posits that experiential leadership programs can "comprehensively transform executives to a higher level of leadership skill and understanding" (299). Yet, like Useem et al. (2005), Judge (2005) cannot empirically validate the claims of experiential leader development effectiveness.

Program Evaluation

It is well understood that evaluation of development programs is important (Alliger, Tannenbaum, Bennett, Traver, & Shotland., 1997; Holton, 1996; Kirkpatrick, 1996; Kirkpatrick, 1998). Evaluation must be "psychometrically sound, meaningful to decision makers, and must be able to be collected within typical organizational constraints" in order to be valuable (Alliger et al., 1997: 342). The most widely used framework for program evaluation is the Kirkpatrick model (Alliger et al., 1997; Arthur, Bennett, Edens, & Bell, 2003; Kirkpatrick, 1998). The Kirkpatrick model provides a taxonomy for identifying four distinct levels in which training transfer can be evaluated: reactions, learning, behaviors, and results (Table 2). By simply and systematically distinguishing between labeled levels of evaluation, the Kirkpatrick model provided the business and academic communities with a valuable tool for promoting the practice and understanding of program evaluation (Alliger et al., 1997).

Table 2.

Kirkpatrick model levels

Level	Evaluation Type	Description & Characteristics
1	Reactions	Participant feelings towards the experience
2	Learning	Knowledge increased through the experience
3	Behaviors	Extent of applied learning taken from the experience
4	Results	Effects on the business or environment realized by the participant due to the experience

The simplicity of Kirkpatrick model's led to its widespread adoption and use (Alliger et al. 1997; Kirkpatrick, 1996). However, the simplicity with which it defines the

levels is also a liability (Alliger & Janak, 1989; Alliger et al., 1997). In response to the perceived shortcomings in the Kirkpatrick model's taxonomy, some researchers developed alternate evaluation models (Alliger et al., 1997; Collins & Holton, 2004; Holton, 1996; Swanson & Holton, 1999). Rather than introduce a new model, however, Alliger et al. (1997) propose an augmented model that more descriptively classifies the levels of learning and offers more specific opportunities for measurement and evaluation (Table 3).

Table 3.

Comparison of Kirkpatrick's and Alliger et al.'s model levels

Level	Kirkpatrick model evaluation type	Alliger et al. (1997) evaluation type
1	Reactions	Reactions a. Affective reactions b. Utility judgments
2	Learning	Learning a. Immediate knowledge b. Knowledge retention c. Behavioral/Skill demonstration
3	Behaviors	Behaviors
4	Results	Results

In Kirkpatrick's (1996) model, level 1 evaluations capture all the participants' feelings toward the training, such as personal enjoyment experienced during the activity, perceived applicability of the activity, and satisfaction with the activity. This focus has led to the labeling of such evaluation instruments as reactionnaires or happy sheets (Hattie et al., 1997; Kirkpatrick, 1998; Lee & Pershing, 2002). Evaluations of participant reactions are typically taken from single source, self-reports (Arthur et al., 2003; Holton,

1996; Kirkpatrick 1996; Hattie et al., 1997; Neill & Richards, 1998). As noted by Kirkpatrick (1996), the results of the level 1 evaluations are often used to guide program development.

Unfortunately, research shows self-reported reactions to be poor indicators of program outcomes and, therefore, a poor guide for program development (Arthur et al., 2003; Kirkpatrick, 1998; Neill & Richards, 1998). Indeed, it is not possible to infer a change in knowledge, skills, or behaviors from self-reported affective reaction results (Alliger et al., 1997; Kirkpatrick, 1998). However, the delineation of affective reactions from utility judgments in Alliger et al.'s (1997) model provides for an increase in evaluative power. Alliger et al. (1997) found that the items used to assess utility judgments often had a more specific focus and produced ratings that correlated with on-the-job performance more highly than those that captured affective reactions. Thus, within level 1 evaluations, utility judgments produce the more useful information.

While level 1 evaluations reveal the reactions of participants to a program, level 2 evaluations focus on capturing the amount of learning experienced by the program participants (Kirkpatrick, 1996). Typically, level 2 evaluations involve the creation and administration of pre- and posttest objective instruments or the observation of participants by trained, third-party observers (Kirkpatrick, 1996; Hattie et al., 1997). Again, Alliger et al. (1997) suggest subdividing the level. By splitting level 2 into three sub-levels: immediate knowledge, knowledge retention, and behavior/skill demonstration, Alliger et al. (1997) distinguish between knowledge measured immediately after instruction, knowledge measured after some time interval within the

program, and behaviors and skills exhibited in the program (Alliger et al. 1997; Craig, 2002).

This delineation is significant as it provides the opportunity to measure the behavioral impact of a program absent the environmental effects that may moderate behavioral change in the job setting (Arthur et al., 2003). As Katz (1956) suggests, it is important that a program participant want to improve, recognize weaknesses, work in a permissive environment, and have the opportunity to try out new ideas in order for behavioral change to occur (as cited in Kirkpatrick, 1996). Using Alliger et al.'s (1997) model for evaluation offers the opportunity to judge the learning facilitated by a program through traditional pre- and posttest knowledge measurements while also capturing the behavioral effect of training in a pure and permissive environment.

The traditional measurement of behavioral change, referred to as training transfer, occurs through on-the-job evaluation in level 3 of Kirkpatrick's model (Kirkpatrick, 1996). It involves pre- and post-program self-reports and pre- and post-program observations provided by bosses, peers, and subordinates (Kirkpatrick, 1998). The more robust the observation pool is, the greater the accuracy is in identifying training transfer (Kirkpatrick, 1996). In addition to the significant measurement and personnel demands necessary to conduct level 3 evaluations, Kirkpatrick (1996) also stresses the necessity of allowing enough time between the program and behavioral evaluations in order to allow changes to occur.

Level 4 evaluations seek to identify the true benefit of the program by tying program outcomes to organizational achievement (Kirkpatrick, 1996). Kirkpatrick (1996) concedes to the difficulty of proving direct complete correlation between program

participation and an organizational achievement when he advises trainers to "be satisfied with the evidence if absolute proof isn't possible to attain" (65-66). Kirkpatrick (1998) semantically draws the distinction between evidence and proof, defining proof as the absolute existence of a causal relationship and evidence as the suggestion of such a relationship.

The difficulty in providing evidence or proof of the relationship between the training and a person's or organization's development is not limited to level 4 but permeates each level of the Kirkpatrick model. Program evaluation becomes more "difficult, complicated, and expensive as it progress from level 1 to level 4—and more important and more meaningful" (Kirkpatrick, 1996: 56). Choosing the correct level of evaluation is a trade-off between the costs of the evaluation and the potential benefits of the measurement (Kirkpatrick, 1998). As noted by Saari, Johnson, McLaughlin, & Zimmerle (1988), the majority of organizations determined that the correct evaluation levels is level 2 or below.

While the Kirkpatrick model enjoys prominence in the evaluation field, Kirkpatrick (1998) stresses the importance of customizing evaluations to the program being evaluated. He borrows a definition of management from the Society for Advancement of Management to define evaluation as both "an art and a science. As a science, it is organized knowledge—concepts, theories, principles, and techniques. As an art, it is the application of organized knowledge to realities in a situation, usually with blend or compromise, to obtain desired practical results" (as cited in Kirkpatrick, 1998: 70). Thus, not only must organizations have the knowledge necessary to create a proper program, they must also possess the ability to design the necessary evaluation.

Evaluating Leader Development

The absence of evaluations of experiential leader development programs beyond the reaction level is not surprising given the fractured nature of the field of leadership, defined to include leadership theory, application, and evaluation. Judge (2005) cites a significant rift between the academic and practitioner communities, with each considering the other largely insignificant to advancing the field.

The majority of the leadership community agrees that leader development is best addressed through a systems approach in which individuals are exposed to developmental opportunities through experience, mentorship, and formal training (Conger & Benjamin, 1999; McCauley & Van Velsor, 2004; Kouzes & Posner, 2002). Though Kouzes & Posner (2002) suggest that formal leader development programs constitute the smallest portion of an effective leadership development mix, they are still significant. The investment in formal leader development programs is substantial, with leader development consuming between 5 and 25 percent of organizational training budgets and costing more than $45 billion annually within the US (Conger & Benjamin, 1999; Kouzes & Posner, 2002). As the level of investment grows, so does the desire to determine the investment's effectiveness.

Evaluating Experiential Learning

The preponderance of research regarding experiential learning programs begins with the assumption that benefits are inherently present in the programs and are designed in a way that does not allow the assumptions to be disproved (Wolfe & Samdahl, 2005). Experiential program evaluations commonly measure participant changes solely by self-reports (Hattie, 1997; Neill & Richards, 1998). Over 80 percent of outdoor (experiential)

education programs limit themselves to the use of level 1 post-program self-report surveys (Neill & Richards, 1998). By using reactions as indicators of program effectiveness, these programs make the subjective impressions of participants the program result of interest. Campbell & Stanley (1963) called such designs "one-shot case stud[ies]" that "have such a total absence of control as to be of almost no scientific value" (6).

Research in experiential program evaluation agrees with Campbell & Stanley's (1963) conclusion. The research reports that the self-reported reactions of program participants have no correlation with program effectiveness and self-reported ratings of effectiveness have no correlation with actual program effects (Neill & Richards, 1998). Thus, experiential programs claims of developmental outcomes are often unsupported by empirical program evaluations.

Those experiential leader development programs that conduct level 2 evaluations often do so through psychometric self-assessment measures in a pretest-posttest design (Hattie et al., 1997; Neill & Richards, 1998; Wolfe & Samdahl, 2005). Such designs allowed for the gathering of tremendous insight into the effectiveness of experiential programs in promoting positive change in self-efficacy and self-perception (Hattie et al., 1997; Neill & Richards, 1998; McKenzie, 2000; McKenzie, 2003; Sheard & Golby, 2006). However, because these evaluations use the Kirkpatrick model and fail to delineate between knowledge and behaviors, such evaluations provide little power in capturing any behavioral changes experienced by program participants.

Evaluating Experiential Leader Development

The evaluation of experiential learning becomes more problematic when done within the context of a leader development program. Of those experiential programs that perform level 2 evaluations, most use generic specific attitudinal measures, not designed to capture specific dimensions of leadership, and instead focus on moderators of leadership, such as self-awareness and locus of control (Hattie, 1997; Neill & Richards, 1998; Sheard & Golby, 2006). A review of program evaluation and leadership literature returned only one study, performed by Keller & Olson (2000), which evaluated the performance of an experiential leader development program against the performance of a non-experiential leader development program. When limited to evaluation of the effectiveness of experiential leader development programs, several studies indicated no perceived growth in leadership (Stolz, 1992). No previous studies or reports in which leadership theories or leadership models served as measurable outcomes of experiential leader development programs could be located.

Problem Statement

This literature review highlights the absence of empirical support for the claimed relationship between experiential learning activities and increases in measured program outcomes, particularly in the field of leader development (Garvin et al., 1996; Keller & Olson, 2000; Roland, 1984; Sheard & Golby, 2006; Wagner et al., 1991; Useem et al., 2005). Yet, support for experiential leader development continues to grow and organizations continue to invest significant resources in such programs without a clear understanding of the value of experiential exercises in leader development (Keller &

Olson, 2000; Williams, Graham, & Baker, 2003). Such is the case with the US Air Force's Squadron Officer School (SOS) and the Combat Leadership Exercise (CLX).

Hypotheses

Using Alliger et al.'s (1997) augmentation of the Kirkpatrick (1996) model, it is possible to evaluate the effectiveness of SOS' leader development program. By evaluating students' leadership before and after SOS, it will be possible to determine if in-residence attendance of SOS program results in leader development.

H1: SOS is positively correlated to leader development.

Comparison of the SOS leader development program's effectiveness before and after the addition of the CLX enables the isolation of the program effect attributable to the CLX. This comparison establishes a means of testing the hypothesis that the curriculum with the CLX will influence leadership development to a greater degree than does the SOS curriculum without the CLX.

H2: The SOS curriculum with the CLX will influence leader development more than does the curriculum without the CLX.

III. Methodology

Sample

Squadron Officer School is a US Air Force professional military education program offered as a five-week in-residence program and as an 18 month distance learning program. It is charged with "develop[ing] dynamic Airmen ready to lead Air, Space, and Cyberspace power in an expeditionary warfighting environment" (http://soc.maxwell.af.mil). To do so, SOS provides instruction in five areas of study: profession of arms, military studies, international studies, communication studies, and leadership and management (http://soc.maxwell.af.mil). The Air Force considers SOS essential to the development of the AF officer corps.

Air Force captains with at least four and fewer than seven years of total active federal commissioned service and Department of Defense (DOD) civilians in the grade of GS-9 and above with at least three years of continuous civil service are eligible for SOS (http://soc.maxwell.af.mil). Completion of the SOS program, either in-residence or via distance learning, is a prerequisite for career advancement for Air Force captains. Approximately 450 eligible members are competitively selected to attend each of the seven in-residence classes held each year.

Those selected for the in-residence class are given temporary duty assignments to Maxwell Air Force Base, Alabama. This assignment places attendees in the employ of SOS and, in so doing, temporarily severs the students from their previous job

responsibilities. Students are expected to focus exclusively on their performance in the SOS program.

Upon their arrival at SOS, students are assigned to student flights of twelve to fifteen people and paired with a trained instructor, the flight commander. The assignment to flights is not random, but is matched to promote demographic homogeneity. The flights are aggregated into four student squadrons, each containing between six and eight flights and headed by a squadron commander. SOS uses this structure to build its program.

SOS segregates its main facility by student squadron with each flight assigned a classroom. The majority of the program's instruction occurs in the classroom and is augmented through guest speakers, experiential exercises, and intramural programs. The instruction schedule fills each weekday from approximately 0700 to 1700. Additionally, SOS requires both individuals and flights to complete work outside the scheduled instruction window. To facilitate gathering for such assignments and encourage social interaction, SOS provides lodging for all in-residence students, organized by flight and squadron.

SOS is an environment of constant instruction and evaluation; performance is subjectively and objectively measured through inputs at the individual, flight, and squadron levels. Students receive subjective evaluation through peer evaluations from each flight member and the flight commander regarding performance in seven key components of leadership. Multiple-choice tests and specified achievements provide objective measurements for each student. The flight commander also provides evaluation at the flight level, subjectively assessing group performance during observed activities

and objectively measuring completion of assigned tasks. The objective flight measurements receive intra-squadron ratings and the top results from each squadron are rated against each other. SOS rewards performance at each level, offering end-of-course individual awards for the top 10% of the students, the top flight, and top squadron as well as weekly flight and squadron awards during the course of the program.

SOS Demographics

The eligibility pool results in a SOS class population with demographics that closely resemble those of the Air Force as whole. While the competitive selection process encourages early attendance for high-performance individuals, the operational demands of the Air Force, the fixed window of opportunity, and moving pool of eligible officers results in SOS classes of mixed high- and mid-range performers.

Not surprisingly, the samples used in this research were largely homogeneous. The comparison group had a mean age of 33.15 years and was predominately male (84.1%) and Caucasian (82.2% Caucasian; 9.3% African-American; 5.6% Asian; 0.9% multi-racial; 1.9% other). The comparison group was well educated (100% Undergraduate, 36.4% Post-Graduate) and served in the Air Force only as an officer (18.7% had prior-enlisted service). Similarly, the treatment group had a mean age of 33.78 years and was predominately male (81.0%) and Caucasian (81.0% Caucasian; 5.1 African-American; 3.8% Asian; 2.5% multi-racial; 7.6% other). Again, the treatment group was well educated (100% Undergraduate, 38.1% Post-Graduate) and served in the Air Force only as an officer (15.5% had prior-enlisted service).

Procedure

This research examines the effect of an experiential exercise on leader development using a pretest/posttest quasi-experimental design with comparison and treatment groups. A comparison of the effect of the SOS leader development program on students attending in-residence classes before and after the introduction of an experiential exercise, the CLX, enables this investigation.

Sample Selection

This research occurred over the course of two consecutive SOS classes. The first class did not participate in the CLX and is the comparison group. The second class participated in the CLX and is the treatment group. The research sampled from each of these two consecutive classes.

As noted, the students in each SOS class are organized into four student squadrons comprised of six to eight student flights. In each of the studied SOS classes, the same student squadron was selected as the sample group. The selected squadron included eight flights (N=107) in the comparison group six flights (N=84) in the treatment group. The squadron selected as the sample in each group is representative of the entire SOS class due to the matched nature of the flights and squadrons.

Measures

This research evaluated the SOS leader development program using Alliger et al.'s (1997) augmentation of the Kirkpatrick model. Due to the complexity and cost associated with obtaining on-the-job evaluations of students from supervisors, peers, and subordinates, a level 2c (behavior/skill demonstration) evaluation was planned.

Leadership

The US Air Force and US Army share the competency-based approach to their leadership models (US Air Force, 2006; US Army, 2006). In order to measure the development of tactical leaders, the US Army Research Institute for the Behavioral and Social Sciences Leader Development Research Unit (ARI LDRU) and the Center for Army Leadership developed the Leader AZIMUTH Check (LAC), a leadership assessment instrument (Appendix A). The LAC uses self and peer evaluations of leadership behaviors to quantify an individual's strengths and weaknesses (Karrasch & Halpin, 1999).

The LAC evaluates leadership behaviors using 72 items distributed along 13 scales, with each scale representing a dimension of leadership. The instrument asks the respondent to indicate how well he thinks each item "describes the person being evaluated" compared with "others [he] has known well" (ARI LDRU, 1998: 3). Responses to each item are recorded on a Likert scale (1 = "Extremely poor description" to 6 = "Extremely good description," with an option of 0 = "Not observed") to indicate the degree to which each statement describes the person being evaluated. Administrations to 42,000 US Army soldiers established the LAC's scale reliability for both self and peer ratings (Table 4).

Table 4.

Leadership Azimuth Check scales and reliabilities

Factor	Scale	Cronbach's α (Self)	Cronbach's α (Peer)
Leadership	Overall	0.97	0.98
Transactional	Decision Making	0.67	0.80
	Planning	0.71	0.78
	Executing	0.84	0.90
	Assessing	0.81	0.86
	Communicating	0.71	0.84
Transformational	Motivating	0.82	0.89
	Building	0.65	0.77
	Developing	0.69	0.81
	Learning	0.85	0.89
Personality/Charisma	Respect	0.76	0.80
	Integrity	0.68	0.75
	Service	0.68	0.73
	Stability	0.82	0.86

Note: Self N =12,660, Peer N = 37,814. (Steele, 2007)

Analysis by ARI of the LAC results produced a second-order path network diagram in which the 13 scales load on three factors: transformational, transactional, and personality/charisma (Steele, 2007). These three first-order factors then load onto the second-order factor of leadership. The Decision Making, Planning, Executing, Assessing, and Communicating scales comprise the transactional factor. The transformational factor

the Motivating, Building, Developing, and Learning scales. The personality/charisma factor holds the Respect, Integrity, Service, and Stability scales. The path network diagram for this model is presented in Appendix B.

The similarities in Air Force and Army leadership development models allow portability of the instrument between the services (US Air Force, 2006; US Army, 2006). However, three items were too specific to the Army and required modification before use in evaluation of the SOS leader development program (Appendix C). The modifications replaced Army-specific terminology with language familiar to the Air Force and did not alter the intent of the items. Additionally, the existing layout of the LAC required adjustment to facilitate self- and peer-report data gathering within the research constraints. To this end, the LAC layout was modified to adjust the Likert scale (1 = "Not at all" to 5 = "To a great extent") and allow for self- and peer-reports to be recorded on one instrument (Appendix D).

Instrument Administration

The AF-specific modified LAC was administered to the samples from the comparison and treatment groups. As administered, the instrument required self- and peer-reports. The self-report required completion of the 72 modified LAC items. To capture peer-reports, each student was randomly assigned three flight members to rate according to the LAC. Ratee assignment was maintained through the pretest and posttest to ensure internal validity was not affected by instrumentation bias introduced by using different raters in instrument administrations.

SOS flight commanders administered paper versions of the modified LAC to the flights in each sample group at similar points in the SOS program. The comparison group

received the pretest on academic day 8; the treatment group completed it on academic day 7. The treatment group completed the CLX on academic day 11 and accomplished guided reflection on the experiential exercise on academic day 14. Posttest administration occurred on academic days 17 and 18 for the comparison and treatment samples, respectively.

In the time before and between administrations, both the comparison and treatment groups completed similar instruction modules and activities. Importantly, each group experienced the same leadership modules (Appendix E). The replacement of Academic Test 2 by the CLX for the treatment group represented the only substantive difference between the comparison and treatment groups' schedules.

Analysis

A traditional pretest/posttest evaluation requires the computation of the means for measured items or scales on both the pretest and posttest (Campbell & Stanley, 1963; Collins, 2002; Collins & Holton, 2004). The pretest mean is then subtracted from the posttest mean to find the difference in means. The mean difference represents the change in the observed measure experienced between the two measures (Campbell & Stanley, 1963; Collins, 2002; Collins & Holton, 2004). Such analysis is common in program evaluation (Collins, 2002; Collins & Holton, 2004; Priest, 2001).

Common statistical applications, such as the Statistical Package for the Social Sciences (SPSS), can evaluate mean differences. After loading the results of the modified LAC into SPSS, a repeated measures ANOVA can evaluation the samples for within group differences across occasions (pretest and posttest) and between group (comparison

and treatment) differences. In traditional analysis, the resultant mean differences and associated effect sizes in the within group across occasions analysis would quantify the ability of the associated SOS curriculum to influence leader development. The between group analysis would establish whether differences between the comparison and treatment groups, and, thereby, their respective curricula, were of statistical significance. Thus, repeated measures ANOVA theoretically can evaluate of the effectiveness of the SOS leader development program for each sample group and determine whether the addition of the CLX influences leader development.

However, the use of mean differences in program evaluation can be problematic (Howard & Dailey, 1979; Howard, Ralph, Gulanick, Maxwell, Nance, & Gerber, 1979; Martineau & Hannum, 2004; Rohs, 1999). To be valid, the posttest-pretest design requires the presence of a common metric in both tests (Cronbach & Furnby, 1970). Metric inequivalence, first described as instrument decay, is a threat to internal validity (Campbell & Stanley, 1963; Craig, 2002). An inherent assumption in posttest-pretest evaluations is the presence of a standard metric between administrations. If the standard of measurement is not constant between administrations, the mean difference will be distorted by the metric inequivalence, thereby stripping the value's validity as a measure of program effectiveness (Campbell & Stanley, 1963; Craig, 2002; Rohs, 1999).

Response Shift

Howard & Dailey (1979) investigated this phenomenon in self-reports, labeling it "response shift." To overcome the effects of response shift, Howard et al. (1979) suggested the use of a retrospective pretest, or then-test. Howard et al. (1979) operationalized response shift as the difference between pretest and then-test

measurements. The then-test, administered after training completion, requires respondents to rate how they believe they were prior to starting the program (Craig, 2002; Rohs, 1999). The then-test avoids the problem of metric inequivalence through temporal proximity to the posttest, which ensures the use of the same perspective in both evaluations. Research indicates that the difference between the posttest and then-test presents a better evaluation of program effectiveness than does conventional posttest-pretest comparison (Craig, 2002; Howard et al., 1979; Rohs, 1999; Sprangers & Schwartz, 1999).

A common consequence in leader development programs is the change in participants' understanding of leadership (Martineau & Hannum, 2004; Rohs, 1999). This change in understanding creates metric inequivalence between pre-program and post-program assessments, resulting in a response shift in posttest-pretest designs. Thus, the then-test appears particularly applicable to the evaluation of leader development (Hannum, Martineau, Reinelt, 2007; Martineau & Hannum, 2004; Rohs, 1999). However, the validity of the then-test is threatened by recall bias (Schwartz, Sprangers, Carey, & Reed, 2004; Visser, Oort, & Sprangers, 2005) and the argument used to establish the utility of the then-test is flawed for,

> If data collected pre-intervention are not on a comparable metric with data collect post-intervention than arithmetic differences between pre and post data are meaningless. Yet, Howard and his colleagues [Howard et al., 1979] operationalize response shift bias as a significant difference between the conventional pretest (collected pre-intervention) and the retrospective pretest (collected post-intervention). (Craig, 2002: 13)

Golembiewski, Billingsley, and Yeager (1976) suggested allowing for multiple changes due to interventions (Craig, 2002). Golembiewski et al. (1976) classified the

types of potential changes as gamma change, beta change, and alpha change. It is possible for differences between observations to be a result of any of the types of change or a combination of the types (Golembiewski et al., 1976). Thus, it is necessary to evaluate observed differences for each type of change (Golembiewski et al., 1976; Craig, 2002).

Gamma Change

Golembiewski et al. (1976) defined change in underlying content domain as gamma change (Craig, 2002). This type of change occurs when respondents reconceptualize the measured domain, giving it a qualitatively different definition between measurements (Craig, 2002). As noted by Craig (2002), Golembiewski et al. (1976) use the example of the change in meaning of "freedom" to African-Americans precipitated by the civil rights movement. Prior to the civil rights movement, "freedom" included the ability to travel by bus, but being relegated to the segregated section in the rear of the bus. After the civil rights movement, "freedom" in travelling meant unrestricted access to the transit system. This simple example illustrates how conceptualization may change over time.

Oort (2005) further divides gamma change into two sub-types: reconceptualization and reprioritization. Oort's (2005) definition of reconceptualization is in line with Golembiewski et al.'s (1976) definition of gamma change. However, Oort offers reprioritization as means to more specifically identify the changes in relative importance a respondent may assign to components within an construct (Oort, 2005). To better illustrate reprioritization, Oort (2005) offers the example of how mental health may

become more important in conceptualization of quality of life than physical health if the patient has a disease that causes significant physical impairment.

Though the tools necessary to identify gamma change were not readily available to Golembiewski et al. (1976), powerful tools are now widely available (Craig, 2002; Oort, 2005). It is possible to use factor analysis to identify the presence of gamma change (Craig, 2002; Oort, 2005; Oort et al., 2005). By contrasting the factor structure and loading of responses before and after an intervention, it is possible to detect whether a response shift due to gamma change, represented by reconceptualization and reprioritization, did occur (Craig, 2002; Oort, 2005; Oort et al., 2005).

Beta Change

Golembiewski et al.'s (1976) definition of beta change parallels that of Howard et al.'s (1979) initial definition of response shift (Craig, 2002; Oort, 2005). According to Golembiewski et al. (1976), beta change represents a recalibration of the metric between measurements. This recalibration causes a respondent to change their interpretations of scale values or labels or both (Oort, 2005; Sprangers & Schwartz, 1999). For example, over the course of treatment, a cancer patient may come to interpret a scale of pain differently, thereby creating a metric inequivalence with longitudinal self-reported pain levels.

Significantly, because beta change accounts for a shift along a common metric, Golembiewski et al. (1976) and Craig (2002) suggests that it can be positively identified only if gamma change is not present. The necessity of hierarchical dependency appears reasonable because changes in internal standards will lose meaning if the underlying construct upon which the standards are based changes as well (Sprangers & Schwartz,

1999). However, Oort's (2005) division of gamma change into the subtypes of reconceptualization and reprioritization should allow for investigation of the lower levels of change as long as gamma change is limited to reprioritization; the composition of the underlying construct does not change, only the relative importance of its components do.

Alpha Change

Once gamma change and beta change are eliminated, the effects of response shift are controlled and true change, or alpha change, can be measured (Craig, 2002; Golembiewski et al., 1976; Oort, 2005). Alpha change is defined as that change found in a respondent's level for a target construct (Oort, 2005) or, more clearly, as the "change from pretest to posttest corresponding to an actual or absolute change" (Millsap & Hartog, 1988: 547). Manifest alpha change often is the goal of development programs, while occurrences of gamma and beta change are frequently treated as measurement errors (Craig, 2002).

Structural Equation Modeling

Oort (2005) and Oort et al. (2005) present an application of structural equation modeling (SEM) techniques that is capable of accounting for each of the three types of change. The use of SEM is significant because it takes a confirmatory, rather than exploratory, approach toward the data analysis and better accounts for variance than do traditional multivariate techniques (Byrne, 2001). Most importantly, SEM can account for both the observed and the unobserved variables (Byrne, 2001).

Oort (2005) and Oort et al. (2005) suggest starting the analysis process with an existing structural equation model, either wholly proven by CFA or modified in accordance with CFA results. For purposes of identification, the initial model should be

constrained such that the common factor means are 0 across occasions (Oort, 2005). Other aspects of the model are free to adjust across measurement occasions, thereby allowing each group to behave independently (Oort, 2005).

The first analytical step is to assess the goodness of fit of the proposed model against the data. The primary test for goodness of fit is the X^2 test. This test evaluates the exact fit of the model using a X^2 value according to a calculated degrees of freedom (Byrne, 2001; Oort, 2005). Because of the improbability of producing an exact model and problems with the robustness of the X^2 test, the X^2 goodness of fit test is often augmented (Byrne, 2001; Oort, 2005). Common measures associated with structural equation modeling goodness of fit tests are the Goodness of Fit Index (GFI), Adjusted GFI (AGFI), Comparative Fit Index (CFI), Parsimony-adjusted CFI (PCFI), and root mean squares error of approximation (RMSEA) (Byrne, 2001). The use of RMSEA is particularly useful as it tests the null hypothesis of a close fit between the model and data (Byrne, 2001; Oort, 2005).

Accepted results for the X^2 test produce $p > .05$, with higher values representing a better fitting model (Byrne, 2001). Results for GFI, AGFI, and CFI are generally accepted when the indices produce results close to 1 (Byrne, 2001). Specifically, the CFI originally was considered to be representative of a well-fitting model with a value > .90 but a revised cut-off of .95 has been recommended (Byrne, 2001). It is reasonable that parsimony-adjusted indices for models exhibiting X^2 statistics and GFI indices in the .90s return values in the .50s, indicating that CFI indices nearing .50 or better are acceptable (Byrne, 2001). The RMSEA test should return values less than .05 for a good fit and as high as .08 for an acceptable fit (Byrne, 2001).

If the goodness of fit test results are acceptable, the hypothesized model serves as the baseline for the analysis of gamma, beta, and alpha change (Model 1). From this baseline model, it is necessary to produce models to evaluate between and within groups. To evaluate the model between the comparison and treatment groups, one additional model must be produced: an invariant model (Model 2). The within group across occasion analysis requires two additional models: an invariant, or no response shift model (Model 2), and a response shift model (Model 3).

In both the between and within group analysis, the invariant model fixes the common factor loads (Γ), intercept means (τ), and variances as constant among the measured groups. In the between group analysis, the evaluation of fits of Models 1 and 2 allow for the determination of comparison group and treatment group equality. For the within group across occasion analysis, differences in fits between Models 1 and 2 provide insight as to whether response shift occurred. Oort et al. (2005) suggest that if model fit does not decrease within groups across occasions (i.e. Model 1 fit is not greater than Model 2 fit within a group), response shift did not occur and analysis may cease.

However, Oort et al. (2005) conducted a one-way comparison of models, using a single group's pretest and posttest scores. The 2x2 structure of this research requires the baseline model (Model 1) to operate between groups and within groups across occasions to conclude that no response shift occurred. Thus, it is necessary to not only compare the fit of Models 1 and 2 at the time of the pretest but also at the time of the posttest to conclude an absence of response shift.

The response shift model emerges in within group across occasion analysis if the relaxation of selected constraints found in Model 2 yields an improved fit. Model 3

develops through an iterative process of constraint analysis and model fitting based on Model 2 (Oort et al., 2005). The process of building Model 3 may be guided by model modification indices and residuals (Byrne, 2001; Millsap & Hartog, 1988; Oort et al., 2005). Once Model 3 is complete, analyzing the differences in fit between Models 1 and 3 allows for the isolation and interpretation of gamma, beta, and alpha changes.

Detecting Gamma Change

Because it is the first in the hierarchical order of changes, gamma change must be addressed first. As noted, gamma change is operationalized as a change in common factor pattern or loading, enabling both reconceptualization and reprioritzation to be assessed through factor analysis through model fit within groups across occasions (pretest, posttest). Determination of reconceptualization requires analysis of the patterns of common factors in the measured group across occasions (pretest, posttest) (Sprangers & Schwartz, 1999; Oort, 2005; Oort et al., 2005). If the pattern of common factors does not change (i.e. no observed variables change the factors on which they load), it is accepted that reconceptualization did not occur between pretest and posttest administrations (Sprangers & Schwartz, 1999; Oort, 2005). Change in common factor loads (Γ) across occasions (pretest, posttest) reveals the presence of reprioritization (Oort et al., 2005; Sprangers & Schwartz, 1999).

Detecting Beta Change

In the context of program evaluation, changes of interest are at the macro, rather than micro, level. Oort (2005) and Oort et al. (2005) define beta change as either uniform or non-uniform recalibration. Because uniform recalibration is representative of beta change for the group (Oort, 2005), investigation of beta change as manifested by uniform

recalibration, rather than non-uniform recalibration, is appropriate for program evaluation.

Within a structural equation model, uniform recalibration is represented by a change in intercept mean (τ) values for the observed variables (Oort, 2005; Oort et al., 2005; Sprangers & Schwartz, 1999). Accordingly, evaluation of the intercept means allows for an assessment of group-level beta change (Oort, 2005; Oort et al., 2005).

Detecting Alpha Change

Oort (2005) operationalizes the measurement instrument's target constructs as the common factors in the model. Thus, changes in common factors can be calculated by comparing common factor means (α) across occasions (Oort, 2005).

Assessing Gamma, Beta, and Alpha Changes

When all change types are identified, it is possible to use X^2 difference tests to assess the statistical significance of the gamma change, beta change, and alpha change. Additionally, the changes can be evaluated for their size and their effect on observed change. Oort (2005) proposes $\mu_2 - \mu_1$ as the model for observed change. Defined further observed change is,

$$\mu_2 - \mu_1 = (\tau_2 - \tau_1) + (\Gamma_2 - \Gamma_1)\alpha_2 + \Gamma_1\alpha_2,$$

where $(\tau_2 - \tau_1)$ represents beta change, $(\Gamma_2 - \Gamma_1)\alpha_2$ constitutes gamma change, and $\Gamma_1\alpha_2$ indicates alpha change (Oort, 2005).

The effect (*d*) of the changes can be found by dividing the observed change and its components by the estimated standard deviation (Oort, 2005). The resultant effect sizes then can be evaluated according to the scale in which $d = 0.2$, 0.5, and 0.8 are considered small, medium, and large effects (Cohen, 1988).

IV. Results

The initially hypothesized model for the LAC provided by ARI (Appendix B) did not fit the data ($X^2_{130} = 346.84$, $p = .000$; RMSEA = .10). The data would not fit any model which included the scales of the Personality/Charisma factor (Respect, Integrity, Service, and Stability). To allow for equivalent evaluation of the data through both mean differences and structural equation modeling, the Personality/Charisma factor was excluded from analysis.

The first-order model that emerged from the fit analysis contained nine of the original 13 scales: Decision Making (DM), Planning (P), Executing (E), Assessing (A), Communicating (C), Motivating (M), Building (B), Developing (D), and Learning (L). The new model retained the first-order factor structure of the initially hypothesized model, with DM, P, E, A, and C loading on The Transactional factor while M, B, D, and L loaded on the Transformational factor. The analysis of the data is limited to the scales and structure of the fitted model.

Instrument Properties

Self-reports

The modified LAC produced expected psychometric properties for the self-reports. The scales are well correlated with one another and have reasonable scale reliabilities. Table 5 presents the psychometric properties of the scales for the self-report

pretest and posttest administrations. The instrument produced excellent overall scale reliabilities for the pretest (α = .944) and posttest (α = .961).

Table 5.

Modified LAC psychometric properties for self-reports

Occasion	Scale	DM	P	E	A	C	M	B	D	L
Pretest	DM	(.607)								
	P	.740	(.706)							
	E	.678	.721	(.732)						
	A	.634	.572	.648	(.743)					
	C	.690	.667	.668	.686	(.714)				
	M	.684	.650	.671	.698	.785	(.783)			
	B	.680	.547	.573	.642	.639	.741	(.723)		
	D	.677	.652	.675	.673	.741	.752	.677	(.614)	
	L	.608	.553	.556	.626	.599	.580	.609	.670	(.552)
Posttest	DM	(.690)								
	P	.756	(.673)							
	E	.711	.752	(.732)						
	A	.663	.696	.725	(.743)					
	C	.714	.761	.740	.722	(.796)				
	M	.673	.730	.761	.752	.773	(.781)			
	B	.681	.665	.718	.694	.741	.773	(.834)		
	D	.752	.740	.739	.743	.723	.774	.754	(.718)	
	L	.732	.697	.660	.713	.645	.692	.694	.754	(.568)

Note: Cronbach α values are on the diagonal.

Peer-reports

Analysis of the peer-reports was surprising. In order to aggregate peer reports, it is necessary to test for the intraclass correlation coefficient (ICC) (Klein, Bliese, Kozlowski, Dansereau, Gavin, Griffin, Hoffman, James, Yammarino, & Bligh, 2000). Because the research design used a different set of randomly selected raters to assess each subject, the ICC(1, 1) model was used (Shrout & Fleiss, 1979). The ICC(1,1) tests showed significance ($p < .05$) for only four of the nine scales (Decision Making, Assessing, Communicating & Motivating). The lack of significance in the remaining five scales indicates that it is not viable to aggregate the values of their peer reported measures (Klein et al.., 2000). With less than half of the targeted peer-report scales viable for aggregation, peer-reports were excluded from further analysis.

Within Groups Across Occasion Mean Differences

Table 6 contains the within groups across occasion (pretest, posttest) mean differences analysis for the comparison and treatment groups. The mean differences provide an initial impression of the change in observed scales without accounting for the possibility of gamma change or beta change (Oort et al., 2005). Table 4.3 also presents the results of the within group across occasion (pretest, posttest) repeated measures ANOVAs and the associated effect sizes (*d*).

The comparison group presented only three scales with statistically significant changes in mean: Executing ($F(1, 4.313)$, $d = -0.209$), Communicating ($F(1, 4.368)$, $d = .185$), and Building ($F(1, 14.138)$, $d = -0.351$). Because these three leader development scales reveal statistically significant negative change, within group across occasions

(pretest, posttest) repeated measures ANOVA of the comparison group provides no support for H1. Effect sizes for the scales with significant changes indicate that the degradation of behaviors across occasion was small.

The differences in means for the treatment group also were small. Only Developing presented a statistically significant change ($F(1, 3.687)$, $d = 0.213$). The change in Developing was positive, indicating growth among the SOS students in this area. The within group across occasion (pretest, posttest) repeated measures ANOVA results support rejection of H1 for all scales but Developing. However, because Developing shows significance, within group across occasions (pretest, posttest) repeated measures ANOVA of the treatment group again provides weak support for H1. The effect size for Developing indicates that the growth was small.

Between Groups Mean Differences

Between group (comparison, treatment) repeated measures ANOVA indicated that no significant differences between the comparison and treatment groups existed. This suggests that the mean differences realized by the groups between the pretests and posttests are statistically equivalent and there is no difference between the comparison and treatment group outcomes. These results support rejection of H2.

Table 6.

Within group across occasions (pretest, posttest) means, mean differences, repeated measures ANOVA, and effect sizes for the comparison and treatment groups

		Pretest		Posttest		Mean	ANOVA		
Group	Scale	Mean	Std Dev	Mean	Std Dev	Diff	df	F	*d*
Comparison	DM	4.0915	0.51531	4.0444	0.54351	-0.0471	1	0.773	-0.089
	P	4.0330	0.58999	3.9239	0.56203	-0.1091	1	3.686	-0.189
	E	4.1556	0.58384	4.0348	0.57003	-0.1208	1	4.313*	-0.209
	A	4.0027	0.61454	3.8995	0.68647	-0.1032	1	2.363	-0.158
	C	4.0942	0.54431	3.9873	0.61236	-0.1069	1	4.368*	-0.185
	M	4.0947	0.61544	4.0174	0.66541	-0.0773	1	1.498	-0.121
	B	4.3315	0.55607	4.1178	0.65839	-0.2137	1	14.138**	-0.351
	D	4.1025	0.51063	4.0616	0.58416	-0.0409	1	0.652	-0.075
	L	4.0739	0.53347	4.0348	0.54540	-0.0391	1	0.536	-0.072
Treatment	DM	3.9601	0.53385	3.9768	0.49771	0.0167	1	0.070	0.032
	P	3.8691	0.60904	3.8989	0.56787	0.0298	1	0.120	0.051
	E	4.1101	0.57899	3.9803	0.54400	-0.1298	1	3.119	-0.231
	A	3.8661	0.68436	3.8443	0.56672	-0.0219	1	0.076	-0.035
	C	3.9079	0.57770	3.8478	0.59521	-0.0601	1	0.854	-0.102
	M	3.9148	0.66177	3.9470	0.63144	0.0322	1	0.158	0.050
	B	4.0997	0.60874	4.0219	0.60436	-0.0779	1	1.099	-0.128
	D	3.8962	0.58564	4.0109	0.48579	0.1148	1	3.687*	0.213
	L	3.9505	0.51808	4.0033	0.48647	0.0527	1	1.085	0.105

Notes: * $p < .05$; ** $p < .01$. Abbreviations: DM – Decision Making; P – Planning; E – Executing; A – Assessing; C – Communicating; M – Motivating; B – Building; D – Developing; L – Learning.

Structural Equation Modeling

The baseline model (Model 1) is a derivation of the LAC model provided by ARI. Figure 2 presents a detailed path diagram of Model 1. Goodness of fit tests revealed that Model 1 fit the data both between (X^2_{44} = 55.9, p = .107; RMSEA = .038, p = .740) and within (comparison: X^2_{44} = 58.243, p = .062; RMSEA = .040, p = .712; treatment: X^2_{44} = 56.783, p = .094; RMSEA = .042, p = .648) groups at the time of the prestest.

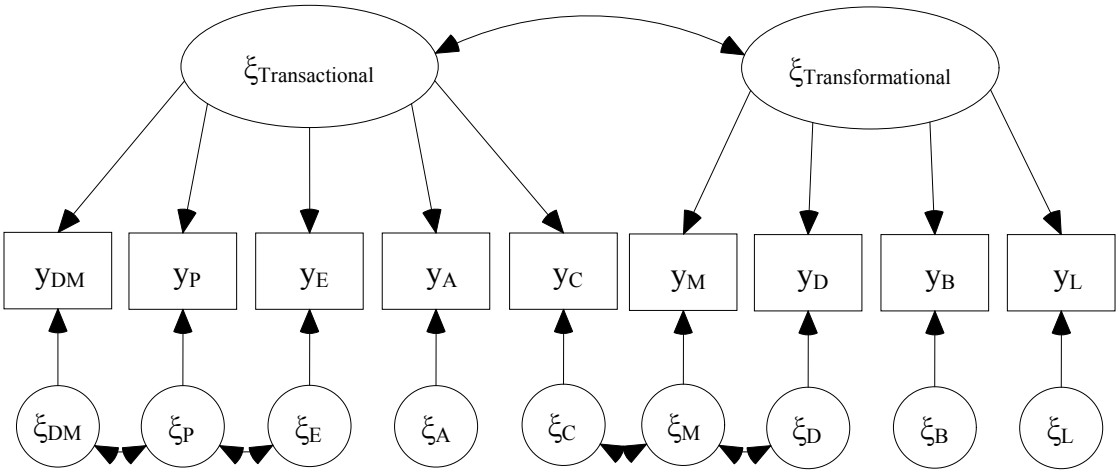


Figure 2. Model 1. Baseline analysis model. *Notes:* Ovals represent latent variables (unobserved common factors), rectangles represent observed variables (modified LAC scales), and circles represent residual factors. Abbreviations: DM – Decision Making; P – Planning; E – Executing; A – Assessing; C – Communicating; M – Motivating; B – Building; D – Developing; L – Learning.

Inter-group Equivalence

In order to establish the equality of the comparison and treatment groups, the fit of Model 1 was compared to the fit Model 2. In Model 2, the invariant model, the common factor loads, intercepts, and variances were constrained between the sample groups. For the purpose of model identification, the common factor means were set at 0 for the comparison group and allowed to freely vary for the treatment group (Oort, 2005). Model 2 yielded a good fit (X^2_{58} = 65.826, p = .224; RMSEA = .027, p = .912). A X^2

difference test between Models 1 and 2 revealed that the comparison and treatment groups were statistically equal at the time of the pretest ($X^2_{14} = 9.926$, p = .768). The equivalence between the comparison and treatment groups at the time of the pretest allows for within group across occasion (pretest, posttest) evaluation using the proposed model.

However, the 2x2 design of this research requires the between group evaluation of models to extend to the time of the posttest. Model 1 produces an adequate fit for the posttest data ($X^2_{44} = 60.097$, p = .054; RMSEA = .044, p = .619). Tests of Model 2 at the time of the posttest show that the model fails to fit the data ($X^2_{58} = 80.664$, p = .026). Model 2's lack of fit between group posttests suggests the groups are inequivalent at the posttest. Therefore, further analysis was necessary to determine the differences between the comparison and treatment groups across occasions.

Intra-group Equivalence

For the comparison group, Model 2 yielded a good fit ($X^2_{58} = 73.870$, p = .078; RMSEA = .036, p = .832). The fit was also good for the treatment group ($X^2_{58} = 69.600$, p = .142; RMSEA = .035, p = .801). In spite of the improved fit experienced in both groups through use of Model 2, the previously identified between group inequivalence at the time of the posttest suggests that each group experienced significant within group changes across occasions, meaning that response shifts did occur.

Comparison Group Response Shift

In accordance with the method used by Oort et al. (2005), the constraints of Model 2 were tested step-by-step to determine each constraint's impact on the within group across occasion model fit for the comparison group. Evaluation revealed that the

best fitting model (X^2_{56} = 62.947, p = .224; RMSEA = .024, p = .944) existed when the intercept means (τ) of the Assessing and Building scales were allowed to freely vary. The model adjusted for these findings, Model 3a (Figure 3), allows for evaluation of gamma, beta, and alpha change. Table 7 presents Model 3a's parameter estimates of interest.

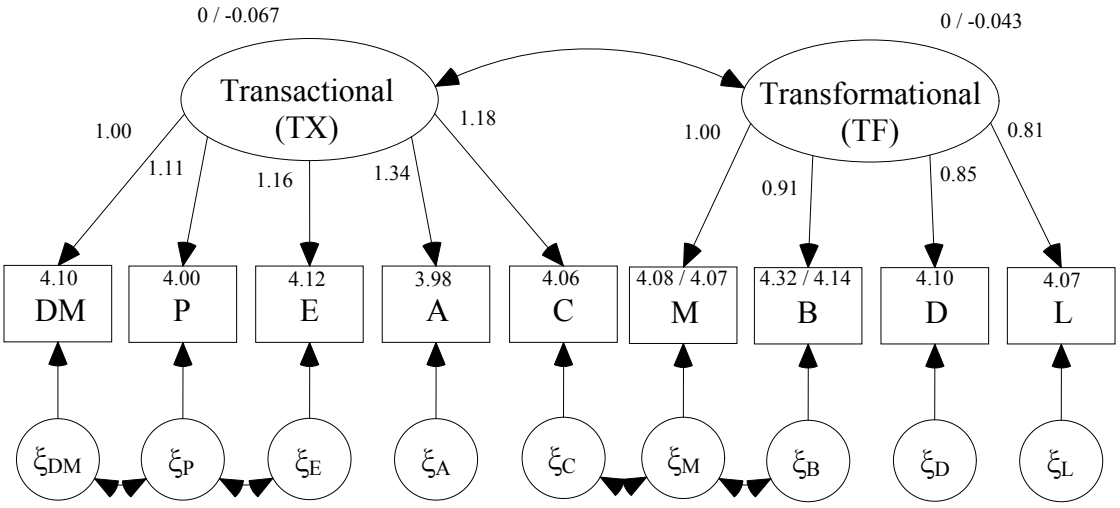


Figure 3. Model 3a. Comparison group response shift model. *Notes:* Parameter estimates separated by a slash indicate pretest and posttest values. Abbreviations: DM – Decision Making; P – Planning; E – Executing; A – Assessing; C – Communicating; M – Motivating; B – Building; D – Developing; L – Learning.

Table 7.

Comparison group parameter estimates in Model 3a.

	Pretest			Posttest	
	TX_1	TF_1		TX_1	TF_2
Factor loadings (Γ)					
DM_1	1.000		DM_2	1.000	
P_1	1.110		P_2	1.110	
E_1	1.159		E_2	1.159	
A_1	1.342		A_2	1.342	
C_1	1.177		C_2	1.177	
M_1		1.000	M_2		1.000
B_1		0.910	B_2		0.910
D_1		0.853	D_2		0.853
L_1		0.806	L_2		0.806

Intercepts (τ)

	DM	P	E	A	C	M	B	D	L
Pretest	4.096	3.996	4.118	3.983	4.061	**4.081**	**4.321**	4.095	4.065
Posttest	4.096	3.996	4.118	3.983	4.061	**4.069**	**4.138**	4.095	4.065

Common factor means (α)

Pretest		Posttest	
TX_1	TF_1	TX_2	TF_2
0.000	0.000	-0.067	-0.043

Notes: n=107, goodness of overall fit measures: $X^2_{56} = 62.947$, RMSEA = 0.244, RMSEA 90% confidence interval = 0.000-0.051. Results indicating significant across-occasion variance are printed in bold ($p < .001$). Though detected, changes in common factor means (α) values are insignificant ($p > .05$). Greek symbols refer to the structural equation model described by Oort (2005). Factor loadings are unstandardized. Abbreviations: DM – Decision Making; P – Planning; E – Executing; A – Assessing; C – Communicating; M – Motivating; B – Building; D – Developing; L – Learning.

Treatment Group Response Shift

The constraints of Model 2 were also evaluated step-by-step using the within group across occasion data of the treatment group. When only the tenable constraints remained, the final model, Model 3b (Figure 4), produced an improved fit ($X^2_{53} = 58.893$, $p = .269$; RMSEA = .026, $p = .882$). The response shift model allowed the factor loadings (Γ) of the Assessing and Building scales and the mean intercepts (τ) of the Decision Making, Building, and Developing scales to vary across the pretest and posttest occasions. Table 8 presents the parameter estimates for Model 3b.

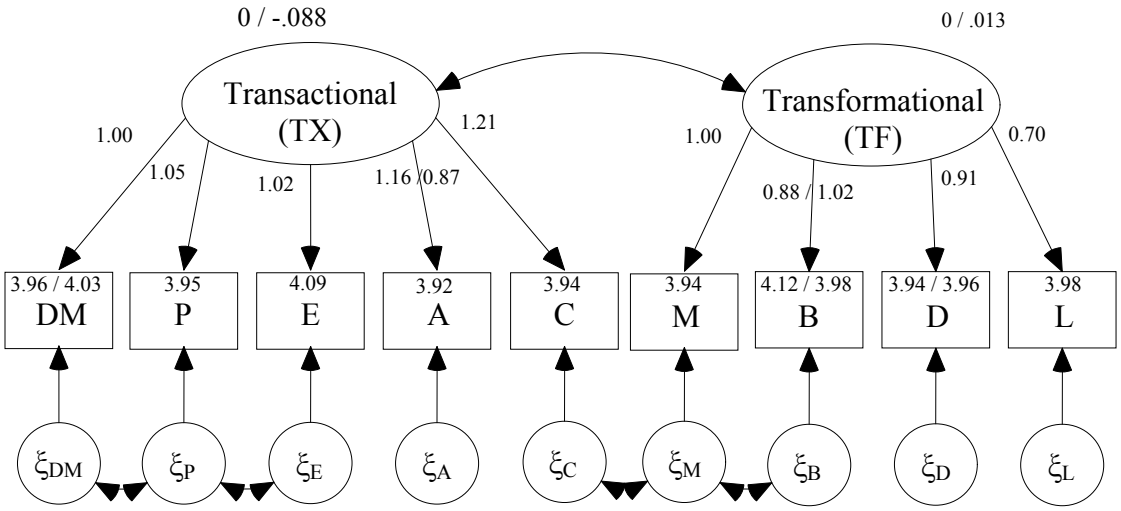


Figure 4. Model 3a. Treatment group response shift model. *Notes:* Parameter estimates separated by a slash indicate pretest and posttest values. Abbreviations: DM – Decision Making; P – Planning; E – Executing; A – Assessing; C – Communicating; M – Motivating; B – Building; D – Developing; L – Learning.

Table 8.

Treatment group parameter estimates in Model 3b.

	Pretest			Posttest	
	TX_1	TF_1		TX_1	TF_2
Factor loadings (Γ)					
DM_1	1.000		DM_2	1.000	
P_1	1.050		P_2	1.050	
E_1	1.018		E_2	1.018	
A_1	**1.160**		A_2	**0.866**	
C_1	1.121		C_2	1.121	
M_1		1.000	M_2		1.000
B_1		**0.883**	B_2		**1.023**
D_1		0.906	D_2		0.906
L_1		0.701	L_2		0.701

Intercepts (τ)

	DM	P	E	A	C	M	B	D	L
Pretest	**3.964**	3.951	4.087	3.918	3.935	3.936	**4.115**	**3.939**	3.980
Posttest	**4.028**	3.951	4.087	3.918	3.935	3.936	**3.980**	**3.963**	3.980

Common factor means (α)

Pretest		Posttest	
TX_1	TF_1	TX_2	TF_2
0.000	0.000	-0.088	0.013

Notes: n=107, goodness of overall fit measures: $X^2_{53} = 58.893$, RMSEA = 0.264, RMSEA 90% confidence interval = 0.000-0.057. Results indicating significant across-occasion variance are printed in bold ($p < .001$). Though detected, changes in common factor means (α) values are insignificant ($p > .05$). Greek symbols refer to the structural equation model described by Oort (2005). Factor loadings are unstandardized. Abbreviations: DM – Decision Making; P – Planning; E – Executing; A – Assessing; C – Communicating; M – Motivating; B – Building; D – Developing; L – Learning.

Evaluating Gamma, Beta, & Alpha Changes Within Groups Across Occasions

In the comparison group, model analysis detected no gamma changes. However, analysis identified significant beta changes in the Motivating and Building scales (Table 9). Thus, SEM analysis found that a response shift, specifically that of recalibration, occurred during the course of the SOS class along these scales. The potential impact of such response shifts on observed scores is highlighted by the effect sizes for the observed change and response shift change on the Building scale. The observed effect size for Building is -0.445, indicating a medium sized negative effect. However, the response shift contribution to the effect size is -0.367. Thus, the students' recalibration of the scale on which they measured Building behaviors accounted for three-quarters of the observed effect. Recalibration of the Motivating scale, though of a proportionally smaller effect size, made a similar impact.

In the treatment group, SEM analysis reveals several significant response shifts (Table 10). The reprioritization subtype of gamma change occurred in the Assessing and Building scales, suggesting that the relative weight students assigned to the respective behaviors changed over the course of the SOS class. Beta change recalibration was found in the Decision Making, Building, and Developing scales. The Building scale again showed the greatest influence of response shift, with reprioritization accounting for 0.004 and recalibration accounting for -0.261 of the -0.236 effect size for observed change. Removing the effects of response shift changes both the magnitude and direction of the change in the Building scale.

Table 9.

Comparison group significance tests of response shifts and effect sizes of observed change, response shift (gamma & beta changes), and true change in Model 3a

	Response Shift	Significance Test		Effect-sizes (*d*)		
Scale		X^2 (df =1)	p	Observed Change	Response Shift	True Change
DM				-0.157		-0.157
P				-0.149		-0.149
E				-0.160		-0.160
A				-0.205		-0.205
C				-0.180		-0.180
M	(τ)Recalibration	9.4	<.001	-0.119	-0.026	-0.093
B	(τ)Recalibration	10.87	<.001	-0.445	-0.367	-0.078
D				-0.098		-0.098
L				-0.081		-0.081

Notes: n = 107; effect-size values of 0.2, 0.5, and 0.8 are considered small, medium, and large (Cohen, 1988). Abbreviations: DM – Decision Making; P – Planning; E – Executing; A – Assessing; C – Communicating; M – Motivating; B – Building; D – Developing; L – Learning.

Table 10.

Treatment group significance tests of response shifts and effect-sizes of observed change, response shift (gamma & beta changes), and true change in Model 3b

	Response Shift	Significance Test		Effect-sizes		
Scale		X^2 (df =1)	p	Observed Change	Response Shift	True Change
DM	(τ)Recalibration		<.001	-0.059	0.158	-0.217
P				-0.184		-0.184
E				-0.208		-0.208
A	(Γ)Reprioritization	2.8	<.001	-0.128	0.043	-0.172
C				-0.228		-0.228
M				0.026		0.026
B	(Γ)Reprioritization	5.3	<.001	-0.236	0.004	0.022
	(τ)Recalibration	9.4	<.001		-0.261	
D	(τ)Recalibration	3.8	<.001	0.107	0.072	0.035
L				0.018		0.018

Notes: n = 84; effect-size values of 0.2, 0.5, and 0.8 are considered small, medium, and large (Cohen, 1988). Abbreviations: DM – Decision Making; P – Planning; E – Executing; A – Assessing; C – Communicating; M – Motivating; B – Building; D – Developing; L – Learning.

V. Discussion

Program Outcomes

While the intent of this research was to evaluate and compare the leader development outcomes of different SOS curriculum, the results of the traditional means difference-based program evaluation proved uninteresting. Such analysis showed little significant change over the course of either SOS class and no significant difference between SOS classes. These rudimentary results suggest that SOS has minimal impact on leader development and that the addition of the CLX experiential exercise made no difference in leader development. However, investigation of the instrument data for response shift delivered interesting results.

The structural equation modeling between groups analysis showed that the comparison and treatment groups were equivalent at the time of the pretest and different at the time of the posttest. This suggests that the differences in SOS curricula, specifically the addition of the CLX, produced different leader development outcomes.

The structural equation modeling analysis within groups across occasions showed that both the comparison and treatment groups registered significant response shifts between the pretest and posttest. The differing response shifts between the comparison and treatment groups again suggest that the addition of the CLX impacted the influence of SOS on leader development. In both groups, the response shifts indicate metric inequivalence between instrument administrations, thus rendering traditional means

difference-based evaluations ineffective as indicators of development and unreliable indicators of program outcomes.

Outcome Effect Sizes

Examination of the program outcomes through structural equation modeling results indicates that the true (α) change effect sizes seen in both groups along the leader development scales are largely trivial, with most changes failing to register even a small effect size. However, the absence of large effect sizes for true change does not indicate a programmatic failure.

Implications

Indeed, Golembiewski et al. (1976) suggested that response shift may be a legitimate outcome of an organizational intervention. Craig (2002) offers the example of how customer service training may have an explicit goal of redefining customer service from being polite and prompt to delivering value for the customer, a type of gamma change. In such a case, a response shift would be a desired outcome.

In discussion of response shift, Martineau & Hannum (2004) indicate that leader development initiative participants "are exposed to a variety of leadership models and to a variety of people whose perspectives about leadership differ from their own. As a result, they leave the initiative with a somewhat different idea of what effective leadership is" (35). For example, an Air Force captain selected for SOS may consider himself to be a good leader, one particularly gifted in team building. Upon entering SOS, he is assigned to a student flight with members with backgrounds and experiences different than his own. In his formal and informal interaction with flight members, he

may find that some members display team building behaviors above or different than his own. Participation with his flight in the activities required by SOS may also cause him to change how he places team building relative to other leader behaviors or change his perceptions regarding the importance of team building in attaining success. Additionally, the formal instruction SOS provides on leadership may cause the captain to adjust how he measures the display of team building behaviors. Thus, though he entered SOS thinking himself to be a good team builder, he may leave with a different idea of what team building is and how it relates to leadership.

In this research, the comparison group changed their perspective as to what it meant to display assessing and building behaviors, developing a more critical perspective over the course of the SOS class. The treatment group members not only became more critical in their perspectives regarding decision making, motivating, and building, but also changed the importance assigned to assessing and building behaviors within the transactional and transformational factors. The response shifts are in line with the SOS mission to "broaden the focus on essential leadership competencies" of its students and may be indicative of program success rather than measurement failure (http://sos.maxwell.af.mil/mission.htm).

Similar to SOS, many leader development programs seek not only to develop KSAs but also to change the way in which participants perceive their worlds and change the way in which they think (Day, 2001). Changes in frames of reference can be "a positive outcome of the program, indicating that participant's knowledge in a particular domain has increased" (Craig & Hannum, 2007: 36). In consideration of such an

outcome, response shifts ought to be measured, not only for separation from true change, but as legitimate outcomes for analysis and evaluation.

Permitting changes in participant perception to be considered legitimate leader development program outcomes changes the leader development program evaluation paradigm. McCauley & Van Velsor (2004) explicitly state that "a key underlying assumption in all our work is that people can learn, grow, and change" (3). While this statement is straight-forward and seemingly obvious, it is difficult to determine best way to capture learning, growth, and change (McCauley & Van Velsor, 2004). In fact, Day & Halpin (2004) noted that "despite the voluminous leadership literature, relatively little is known about what exactly gets developed in leader development" (4-5).

Increase in participant awareness and understanding necessarily precede any behavioral outcomes or organizational results that stem from a development program (Martineau, 2004). As such, it is prudent that program evaluation address not only observed changes but also measure changes in conceptualization and frames of reference. Though the analytical techniques required for structural equation modeling are more advanced than means difference-based evaluations (Craig & Hannum, 2007), structural equation modeling presents the only method that is capable of measuring both observed and latent variables (Byrne, 2001). It allows program evaluation to capture the all the changes achieved in development programs, both in behavior and in thought, by quantifying gamma, beta, and alpha change.

Such an assessment of change and response shifts made possible by structural equation modeling is particularly applicable to experiential leader development programs. Kolb's (1984) ELT explicitly states that participants will experience changes in

conceptual understanding which will translate into future actions. Thus, gamma and beta changes are crucial to Kolb's interpretation of the learning cycle. Evaluations of experiential programs that do not capture response shifts ignore the delivery method's influence on the learning cycle and fail to distinguish the outcomes of experiential learning from those available in traditional pedagogies.

The existing evaluation paradigm suggests that program evaluation must extend beyond within-program changes and incorporate measurable on-the-job behavior changes and organizational results (Phillips & Schmidt, 2004; Swanson & Holton, 1999). Yet, as noted by Katz (1956, as cited in Kirkpatrick, 1996) and Craig & Hannum (2007), the ability to alter behavior or deliver organizational benefits are highly moderated by environmental factors. The ideal evaluation would isolate the program evaluation from environmental factors (Craig & Hannum, 2007). The study presented here measured gamma, beta, and alpha changes within a controlled training environment, providing a pure representation of the program's impact immediate impact. However, the utility of structural equation modeling for program evaluation does not end with program completion.

Quality of life research indicates that structural equation modeling techniques are viable methods for measuring the three types of change across various lengths of time and across multiple interventions (e.g. Ahmed, Mayo, Wood-Dauphinee, Hanley, & Cohen, 2005; Oort, 2005; Oort et al., 2005; Schwartz & Sprangers, 2004). Thus, structural equation modeling may be used across a greater longitudinal space, allowing for immediate and follow-up evaluations of program outcomes. Such application would allow for organizations to evaluate the impact of development programs, both immediate

and over time, as well as assess the impact of the organizational environment on program outcomes.

Limitations

The greatest limitation of this research was in its data collection. The scheduling of SOS classes and the intensive time demands within the SOS program made securing a robust sample difficult and resulted in unequal sample sizes, with a comparison group of n = 107 and a treatment group of n = 84. The usable sample size also decreased across occasions due to respondent mortality as non-response increased from 4.2 percent to 17.3 percent between pretest and posttest administrations. Several respondents indicated they experienced survey fatigue, expressing dissatisfaction with the length of the measurement instrument and the perceived repetition in the administrations.

The student's perceptions of the utility of the instrument also may have impacted the quality of data provided in peer ratings. As previously noted, the peer ratings were unusable due to insignificant results in analysis of interclass correlation coefficients. This forced reliance on self-reports. While the poor correlation of peer reports may be an aberration found only in these samples, future research may be improved by requiring ratings from objective observers, such as the flight commanders, at the times of the instrument administrations.

The reliance of structural equation modeling on a hypothesized model also introduces limitations. If, as in this research, the data does not fit the hypothesized model then the model must be adjusted. In the research, the adjustments to the model were based on the existing and confirmed model provided by ARI (Appendix B). While the

adjustments were guided by statistics, they were ultimately subjective, especially in the assignment of covariance among the residual variances.

Future Research

Most evaluations of leader development programs continue to treat response shifts as measurement errors rather than desired outcomes. In spite of the logical fallacy inherent in the foundation of the then-test, its use is encouraged to attenuate response shift bias (Howard et al., 1979; Hannum, Martineau, & Reinelt, 2007; Martineau & Hannum, 2004; Rohs, 1999). A return to traditional pretest/posttest designs accompanied by structural equation modeling analysis may produce significant findings regarding the true outcomes of leader development programs, especially those programs based on experiential learning theory.

The reliance of structural equation modeling on hypothesized models also presents opportunities for future research. Program design should incorporate delivery and evaluation methods (Baldwin et al., 2004; Martineau & Hannum, 2004). If the program design and delivery are tied to the evaluation method and an evaluation model, structural equation modeling offers a way in which to test the fit of the intended results against actual results. This capability will enable further research into the relationship between program design and deliverable results (Baldwin et al., 2004). Ultimately, such research may lead to the identification of variables that predict the presence or magnitude of response shift. Such insight would allow for better program design, whether response shift is seen as a legitimate outcome or measurement error.

The power of structural equation modeling may be exploited further to bridge the chasm between the practitioner and academic communities within the leadership field. The use of SEM allows the researcher or practitioner to examine the fit of program outcomes and evaluation models to leadership theory, such as those proposed by transformational leadership. By joining the academic and practitioner communities in the evaluation of leader development, structural equation modeling will increase the understanding and advance of the fields of leadership theory and leader development.

References

1. Ahmed, S., Mayo, N. E., Wood-Dauphinee, S., Hanley, J. A., and Cohen, R. (2005). The structural equation modeling technique did not show a response shift, contrary to the results of the then-test and the individualized approaches. *Journal of Clinical Epidemiology*, 58, 1125–1133.

2. Albertson, D. S. (1995). Evaluating experiential training. *Developments in Business Simulation & Experiential Exercises*, 22, 166-171.

3. Alliger, G. M. & Janak, E. A. (1989). Kirkpatrick's levels of training criteria: thirty years later. *Personnel Psychology*, 42(2), 331-342.

4. Alliger, G. M., Tannenbaum, S. I., Bennett, W., Jr., Traver, H., & Shotland, A. (1997). A meta-analysis of the relations among training criteria. *Personnel Psychology*, 50, 341-358.

5. Arthur, W., Jr., Bennett, W., Jr., Edens, P. S., & Bell, S. T. (2003). Effectiveness of training in organizations: a meta-analysis of design and evaluation features. *Journal of Applied Psychology*, 88(2), 234-245.

6. Baldwin, C., Persing, J, & Magnuson, D. (2004). The role of theory, research, and evaluation in adventure education. *Journal of Experiential Education*, 26(3), 167-183.

7. Baker, A. C., Jensen, P. J., & Kolb, D. A. (2002). *Conversational learning: an experiential approach to knowledge creation*. Wesport, CT: Quorum Books.

8. Buller, P. F., Cragun, J. R., & McEvoy, G. M. (1991). Getting the most out of outdoor training. *Training & Development Journal*. March, 58-61.

9. Burke, M. J. & Day, R. R. (1986). A cumulative study of training. *Journal of Applied Psychology*, 71: 232-265.

10. Byrne, B. M. (2001). *Structural equation modeling with AMOS*. Marwah, NJ: Lawrence Erlbaum Associates.

11. Campbell, D. T. & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago: McNally.

12. Collins, D. B. (2002). *The effectiveness of managerial leadership development programs: a meta-analysis of studies from 1982-2001*. Unpublished doctoral dissertation, Louisiana State University, Baton Rouge.

13. Collins, D. B. & Holton, E. F. (2004). The effectiveness of managerial leadership development programs. *Human Resource Development Quarterly*, 15(2), 217-248.

14. Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

15. Conger, J. A., & Benjamin, B. (1999). *Building leaders: How successful companies develop the next generation.* San Francisco: Jossey-Bass.

16. American Productivity & Quality Center (APQC). (2000). "Developing Leaders at All Levels." Consortium Benchmarking Study Best Practice Report. Houston, TX: Author.

17. Craig, S. B. (2002). *Implicit theories and beta change in longitudinal evaluations of training effectiveness: an investigation using item response theory*. Unpublished doctoral dissertation, Virginia Polytechnic Institute and State University.

18. Craig, S. B. & Hannum, K. M. (2007). Experimental and quasi-experimental evaluations. In K. M. Hannum, J. W. Martineau, & C. Reinelt. (2007). *The handbook of leadership development evaluation* (pp. 19-47). San Francisco: Jossey-Bass.

19. Cronbach, L. J., Furby, L. (1970). How should we measure "change"—or should we? *Psychology Bulletin*, 74, 68–80.

20. Day, D. V. (2001). Leadership development: a review in context. *Leadership Quarterly*, 11(4), 581-613.

21. Day, D. V. & Halpin, S. M. (2004). Growing leaders for tomorrow: an introduction. In D. V. Day, S. J. Zaccaro, & S. M. Halpin (Eds.). *Leader development for transforming organizations: growing leaders for tomorrow* (pp. 3-21). Mahwah, NJ: Lawrence Erlbaum Associates, Inc. Publishers.

22. Fiedler, F. E. (1996). Research on leadership selection and training: One view of the future. *Administrative Science Quarterly, 41*, 241–250.

23. Garvin, J. D., Nason, E. R., & Otto, J. T. (1996). Technical Report on the United States Air Force Academy's Leadership Reaction Course. (DTIC Accession No. ADA319727). USAFA, CO: Dean of the Faculty.

24. Golembiewski, R. T., Billingsley, K., & Yeager, S. (1976). Measuring change and persistence in human affairs: types of change generated by OD designs. *Journal of Applied Behavioral Science*, 12 (2), 133-157.

25. Hannum, K., Martineau, J. W., Reinelt, C. (2007). *The handbook of leadership development evaluation*. San Francisco: Jossey-Bass.

26. Hattie, J., Marsh, H. W., Neill, J. T., & Richards, G. E. (1997). Adventure education and Outward Bound. *Review of Educational Research*, 67(1), 43-87.

27. Hernez-Broome, G. & Hughes, R. L. (2004). Leadership development. *Human Resource Planning*, 27, 24-32.

28. Hollenbeck, G. P., McCall, M. W., Jr., & Silzer, R. F. (2006). Leadership competency models. *Leadership Quarterly*, 17, 398-413.

29. Holton, E. F. (1996). The flawed four-level evaluation model. *Human Resource Development Quarterly*, 7(1), 5-21.

30. Howard, G. S., & Dailey, P. R. (1979). Response-shift bias: A source of contamination of self-report measures. *Journal of Applied Psychology*, 64, 144-150.

31. Howard, G. S., Ralph, K. M., Gulanick, N. A., Maxwell, S. E., Nance, D. W., & Gerber, S. K. (1979). Internal validity in pretest-posttest self-report evaluations and a reevaluation of retrospective pretests. *Applied Psychological Measurement*, 3, 1-23.

32. Judge, W. (2005). Adventures in creating an outdoor leadership challenge course for an eMBA program. *Journal of Management Education*, 29(2), 284-300.

33. Karrasch, A. I. & Halpin, S. M. (1999). Feedback on 360 degree Leader AZIMUTH check assessment conducted at Fort Clayton, Panama (ARI Research Note 99-21). Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences.

34. Kayes, D. C. (2002). Experiential learning and its critics: Preserving the role of experience in management learning and education. *Academy of Management Learning and Education,* 1(2), 137-149.

35. Keller, T. & Olson, W. (2004). The advisability of outdoor leadership training. *Review of Business*, Spring, 4-6.

36. Klein, K. J., Bliese, P. D., Kozlowski, S. W. J., Dansereau, F., Gavin, M. B., Griffin, M. A., Hoffman, D. A., James, L. R., Yammarino, F. J., Bligh, M. C. (2000). Multilevel analytical techniques. In Klein, K. J., Kozlowski, S. W. J. (Eds.), Multilevel theory, research, and methods in organizations (pp. 512-553). San Francisco: Josey-Bass.

37. Kirkpatrick, D. (1996). Great Ideas Revisited. *Training & Development*. 50(1), 54-59.

38. Kirkpatrick, D. L. (1998). *Evaluating Training Programs: the four levels* (2nd ed). San Francisco: Berrett-Koehler.

39. Kolb, D. A. (1984). *Experiential learning: Experience as the source of learning and development*. New Jersey: Prentice Hall.

40. Kolb, D. A. & Boyatsiz, R. E. (2000). Experiential learning theory: previous research and new directions. In Sternberg, R.J. & Zhang, L.F. (Eds.) *Perspectives on cognitive, learning, and thinking styles* (pp 227-248). New Jersey: Lawrence Erlbaum.

41. Kouzes, J. M. & Posner, B. Z. (2002). *The leadership challenge* (3rd ed.). San Francisco: Jossey-Bass.

42. Kram, K. E. (1985). *Mentoring at work: developmental relationships in organizational life.* Glenview, IL: Scott Foresman.

43. Kram, K. E., & Isabella, L. A. (1985). Mentoring alternatives: The role of peer relationships in career development. *Academy of Management Journal, 28*, 110–132.

44. Lee, S. H., & Pershing, J. A. (2002). Dimensions and design criteria for developing training reaction evaluations. *Human Resource Development International*. 5(2), 175-197.

45. Lombardo, M., & Eichinger, R. (2002). *The leadership machine*. Minneapolis: Lominger Limited, Inc.

46. Martineau, J. (2004). Evaluating the impact of leader development. In McCauley, C. & Van Velsor, E. (Eds.), *The Center for Creative Leadership handbook of leadership development* (pp. 234-267). San Francisco: Jossey-Bass.

47. Martineau, J., & Hannum, K. (2004). *Evaluating the impact of leadership development*. Greensboro, NC: Center for Creative Leadership.

48. McCall, M. W. (1998). *High flyers: developing the next generation of leaders.* Boston: Harvard Business School.

49. McCall, M. W., Lombardo, M. M., & Morrison, A. M. (1998). *The lessons of experience: how successful executives develop on the job*. Lexington, MA: Lexington Books.

50. McCauley, C. & Van Velsor, E. (Eds.). (2003). *The Center for Creative Leadership handbook of leadership development* (2nd ed.). San Francisco: Jossey-Bass.

51. McKenzie, M. D. (2000). How are adventure education program outcomes achieved? *Australian Journal of Outdoor Education*, 5(1), 19-28.

52. McKenzie, M. D. (2003). Beyond "The Outward Bound Process." *Journal of Experiential Education*, 26(1), 8-23.

53. Miettinen, R. (1998). About the legacy of experiential learning. *Lifelong Learning in Europe, 3*, 165-171.

54. Millsap, R.E., & Hartog, S.B. (1988). Alpha, beta, and gamma change in evaluation research: A structural equation approach. *Journal of Applied Psychology*, 73, 574-584.

55. Mintzberg, H. (2004). *Managers not MBAs*. San Francisco: Berrett-Koehler.

56. Neill, J. T. & Richards, G. E. (1998). Does outdoor education really work? *Australian Journal of Outdoor Education*, 3(1), 1-9.

57. Oort, F. J. (2005). Using structural equation modeling to detect response shifts and true change. *Quality of Life Research*, 14, 587-598.

58. Oort, F. J., Visser, M. R. M., & Sprangers, M. A. G. (2005). An application of structural equation modeling to detect response shifts and true change in quality of life data from cancer patients undergoing invasive surgery. *Quality of Life Research*, 14, 599-609.

59. Phillips, J. J. & Schmidt, L. (2004). *The Leadership Scorecard*. New York: Elsevier.

60. Priest, S. (2001). A program evaluation primer. *The Journal of Experiential Education*, 24(1), 34-40.

61. Raelin, J. A. (2004). Don't bother putting leadership into people. *Academy of Management Executive*, 18(3), 131-135.

62. Rohs, F. R. (1999). Response shift bias: a problem in evaluating leadership development with self-report prestest-postest measures. *Journal of Agricultural Education*, 40(4), 28-37.

63. Roland, C. (1984). Outdoor managerial training programs. The Bradford Institute. 69-77.

64. Ronan, J. (2003). A boot to the system. *T + D*, 57(3), 38-45.

65. Saari, L.M., Johnson, T.R., McLaughlin, S.D., & Zimmerle, D.M. (1988). A survey of management training and education practices in U.S. companies. *Personnel Psychology*, 41,731-743.

66. Schwartz, C. E., Sprangers, M. A. G., Carey, A., & Reed, G. (2004). Exploring response shift in longitudinal data. *Psychology and Health*, 19(1), 51-69.

67. Sheard, M. & Golby, J. (2006). The efficacy of an outdoor adventure education curriculum on selected aspects of positive psychological development. *Journal of Experiential Education*, 29(2), 187-209.

68. Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, 86(2), 420-428.

69. Sprangers, M. A. G. & Schwartz, C. E. (1999). Integrating response shift into health-related quality of life research: a theoretical model. *Social Science & Medicine*, 48, 1507-1515.

70. Steele, J. P. (2007). Basic psychometric properties of AZIMUTH multi-rater 6007. Prepared for Lt Col Kent C. Halverson, USAF. Ft. Leavenworth, KS: Ft. Leavenworth Research Unit, US Army Research Institute for the Behavioral and Social Sciences.

71. Stolz, P. G. (1992). An examination of leadership development in the great outdoors. *Human Resource Development Quarterly*, 3, 357-372.

72. Sullivan, M. & Kolb, D. A. (1995). Perspectives in corporate training and development. In C. Roland, D. Wagner, & R. Weigand (Eds.). *Do it and understand!* (pp. 5-11). Dubuque, IA: Kendall/Hunt Publishing.

73. Swanson, R. A. & Holton, E. F. (1999). *Results: how to assess performance, learning, and perceptions in organizations*. San Francisco: Berrett-Koehler.

74. Wagner, R. J., Baldwin, T. T., & Roland, C. C. (1991). Outdoor training: revolution or fad? *Training and Development Journal*, March, 51-57.

75. Wagner, R. J. & Fahey, D. (1992). An empirical evaluation of a corporate outdoor-based training program. Presented at the *Coalition for Education in Outdoors Research Symposium*, Bradford Woods, Indiana, April 22-25, 1992.

76. Wiegand, R. (1995). Experiential learning: a brief history. In C. Roland, D. Wagner, & R. Weigand (Eds.). *Do it and understand!* (pp. 2-4). Dubuque, IA: Kendall/Hunt Publishing.

77. Williams, S. D., Graham, T. S., & Baker, B. (2003). Evaluating outdoor experiential training for leadership and team building. *The Journal of Management Development*, 22(1/2), 45-59.

78. Wolfe, B. D. & Samdahl, D. M. (2005). Challenging assumptions. *Journal of Experiential Education*, 28(1), 25-43.

79. US Air Force. (2006). *Leadership and Force Development* (AFDD 1-1). Washington, DC: Author.

80. US Army. (2006). *Army Leadership* (FM 6-22). Washington, DC: Author.

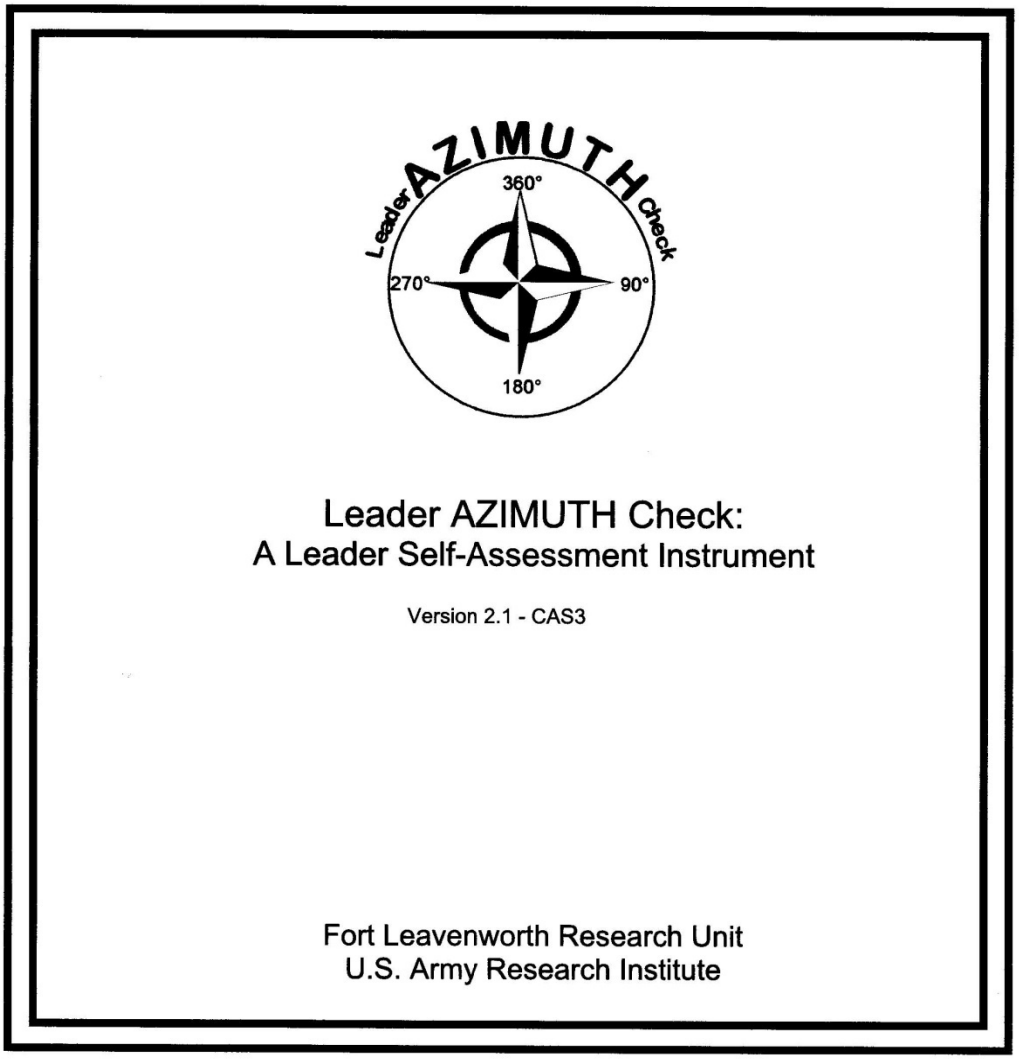
81. US Army Research Institute for Behavioral and Social Sciences Leader Development Research Unit (ARI LDRU). (1998). Leader AZIMUTH Check (PT 60-07). Ft. Leavenworth, KS: Ft. Leavenworth Research Unit, US Army Research Institute for the Behavioral and Social Sciences. For further information contact (913) 684-9753.

82. Useem, M., Davidson, M., & Wittenberg, E. (2005). Leadership development beyond the classroom. *International Journal of Leadership Education*, 1(1), 159-178.

83. Vicere, A. A. & Fulmer, R. M. (1988). *Leadership by design*. Boston, MA: Harvard Business School.

84. Visser, M. R. M., Oort, F. J., Sprangers, M. A. G. (2005). Methods to detect response shift in quality of life data: a convergent validity study. *Quality of Life Research*, 14, 629-639.

Appendix A. Leadership AZIMUTH Check instrument (PT60-07)



PURPOSE: This questionnaire has been designed by the U.S. Army Research Institute and the Center for Army Leadership to obtain information in support of leader self-development. The items in the questionnaire and the feedback based on the items are consistent with current and emerging Army Leadership Doctrine.

CONFIDENTIALITY: The individual ratings and the overall results are provided to the person who is being rated; the information is not provided to anyone in the officer's rating chain. If you are providing ratings on someone else, your input is anonymous.

PT60-07
6Apr'98

Leader AZIMUTH Check

Introduction

The Army places special emphasis on self-development to enhance the leadership skills of military and civilian leaders. As part of self-development, it is important for individuals to become aware of their own strengths and weaknesses. You are asked to provide input on the strengths and weaknesses of the designated officer. AZIMUTH provides each person with feedback based on a comparison of their own self-perceptions and others' perceptions of them. This information is needed from you in order to provide complete and high quality information for the rated individuals. YOUR VOLUNTARY PARTICIPATION IS NEEDED. You are encouraged to answer all questions, but failure to respond to any item will not result in any penalty.

The identification numbers and names on the AZIMUTH answer sheets are provided to identify the person being rated. When you are rating someone else your rating is ANONYMOUS; no record is kept of who rates whom. However, if you do not respond to all the questions, then the person being assessed will receive incomplete feedback. If you are doing a self-assessment, rating yourself, you need to be aware that the self-assessment cannot be anonymous; we need to be able to identify you in order to provide you feedback. Only persons involved in collecting or preparing the information for analysis will have access to completed AZIMUTH forms. Any reports of these data will contain only group statistics.

Instructions

If you are using this form for self-assesment: 1) Be sure to read and sign the Privacy Act Statement before proceeding. 2) Fill in your own name and ID number on all mark-sense forms to be completed by yourself and others. 3) Fill in the bubble at the top of page 3 indicating that the person being rated is your self. 4) Complete one self form by marking the bubbles which best indicate how well each item describes you.

If you are rating a classmate: The classmate you are rating should have already filled in their name and ID number. Please: 1) Skip the Privacy Act Statement section. 2) Fill in a bubble at the top of page 3 to indicate that the person being rated is your classmate. 3) Mark the bubbles which best indicate how well each item describes this classmate.

PRIVACY ACT STATEMENT:

Public Law 93-573, called the Privacy Act of 1974, requires that you be informed of the purpose and uses to be made of any information collected.

The Department of the Army may collect the information requested in this questionnaire under the authority of 10 United States Code 2358. Providing information in this questionnaire is voluntary. Failure to respond to any particular questions will not result in any penalty. However, if you are providing an assessment of yourself, then failure to provide your ID number will prevent you from receiving feedback for your leadership self-development.

The primary use of the information collected will be to provide the person being rated with feedback for his/her leadership self-development. The aggregate data will also be used by the U. S. Army Research Institute for research and development purposes. Your responses will be held in strict confidence. No responses or summaries, whole or in part, will become a part of any individual's personnel file. **This information will not be used by anyone for an evaluation of the person being assessed - it will be used to provide him/her with feedback for self-development.**

(If you are providing an assessment of someone else, then please DO NOT enter your name or signature.)

PRINT your name here: _____ Date: _____

I authorize use of this information as stated above: _____

(Sign Your Name Above)

Name of person being rated:

_____

MARK THIS ITEM FIRST

The Person Being Rated is my:

- ○ Self
- ○ Classmate

Identification Number

0	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2	2
3	3	3	3	3	3	3	3	3
4	4	4	4	4	4	4	4	4
5	5	5	5	5	5	5	5	5
6	6	6	6	6	6	6	6	6
7	7	7	7	7	7	7	7	7
8	8	8	8	8	8	8	8	8
9	9	9	9	9	9	9	9	9

MARKING INSTRUCTIONS

- Use a No. 2 pencil only.
- Do not use ink, ballpoint, or felt tip pens.
- Make solid marks that fill the response completely.
- Erase cleanly any marks you wish to change.
- Make no stray marks on this form.

CORRECT: ● INCORRECT: ✓ ⊗ ◒ ◔

In comparison with others I have known well, I think the items below describe the person being rated as indicated.

	Have Not Observed	Extremely Poor Description	Very Poor Description	Slightly Poor Description	Slightly Good Description	Very Good Description	Extremely Good Description
Communicating							
1. Provides clear direction.	○	○	○	○	○	○	○
2. Explains own ideas so that they are easily understood.	○	○	○	○	○	○	○
3. Keeps others well informed.	○	○	○	○	○	○	○
4. Listens well.	○	○	○	○	○	○	○
5. Tells it like it is.	○	○	○	○	○	○	○
6. Communicates poorly.	○	○	○	○	○	○	○
Decision-Making							
7. Delays decisions unnecessarily.	○	○	○	○	○	○	○
8. Generates innovative solutions to unique problems.	○	○	○	○	○	○	○
9. Ignores information that conflicts with own initial assumptions.	○	○	○	○	○	○	○
10. Makes sound decisions.	○	○	○	○	○	○	○
11. Willing to revisit a decision when new information calls for it.	○	○	○	○	○	○	○
12. Effectively uses CAS3 problem solving techniques.	○	○	○	○	○	○	○
Motivating							
13. Provides good explanation of rationale when directing tasks.	○	○	○	○	○	○	○
14. Contributes to a supportive environment.	○	○	○	○	○	○	○
15. Inspires people to do their best.	○	○	○	○	○	○	○
16. Quick to acknowledge good performance of others.	○	○	○	○	○	○	○
17. Sets clear performance expectations.	○	○	○	○	○	○	○
Developing							
18. Encourages professional growth.	○	○	○	○	○	○	○
19. Is an effective teacher.	○	○	○	○	○	○	○
20. Provides honest feedback to others on their strengths and weaknesses.	○	○	○	○	○	○	○
21. Sets the example for others by doing his or her best.	○	○	○	○	○	○	○
22. Seldom shares responsibility with others.	○	○	○	○	○	○	○
Building							
23. Actively participates in the activities of the staff group.	○	○	○	○	○	○	○
24. Encourages cooperation among staff group members.	○	○	○	○	○	○	○
25. Solicits group input in decision-making situations.	○	○	○	○	○	○	○
26. Focuses the staff group on mission accomplishment.	○	○	○	○	○	○	○
27. Treats others as valuable team members.	○	○	○	○	○	○	○
Learning							
28. Becomes defensive when given critical feedback.	○	○	○	○	○	○	○
29. Encourages open discussion to improve the staff group.	○	○	○	○	○	○	○
30. Helps the staff group adapt to changing circumstances.	○	○	○	○	○	○	○
31. Seems to be realistic about own personal limitations.	○	○	○	○	○	○	○
32. Is willing to accept new challenges.	○	○	○	○	○	○	○
Planning and Organizing							
33. Anticipates how different plans will look when executed.	○	○	○	○	○	○	○
34. Develops effective plans to achieve staff group goals.	○	○	○	○	○	○	○
35. Leaves key events to chance.	○	○	○	○	○	○	○
36. Sets clear priorities.	○	○	○	○	○	○	○
37. Is unwilling to modify original plan when circumstances change.	○	○	○	○	○	○	○
38. Manages time effectively.	○	○	○	○	○	○	○

PT60-07

In comparison with others I have known well, I think the items below describe the person being rated as indicated.

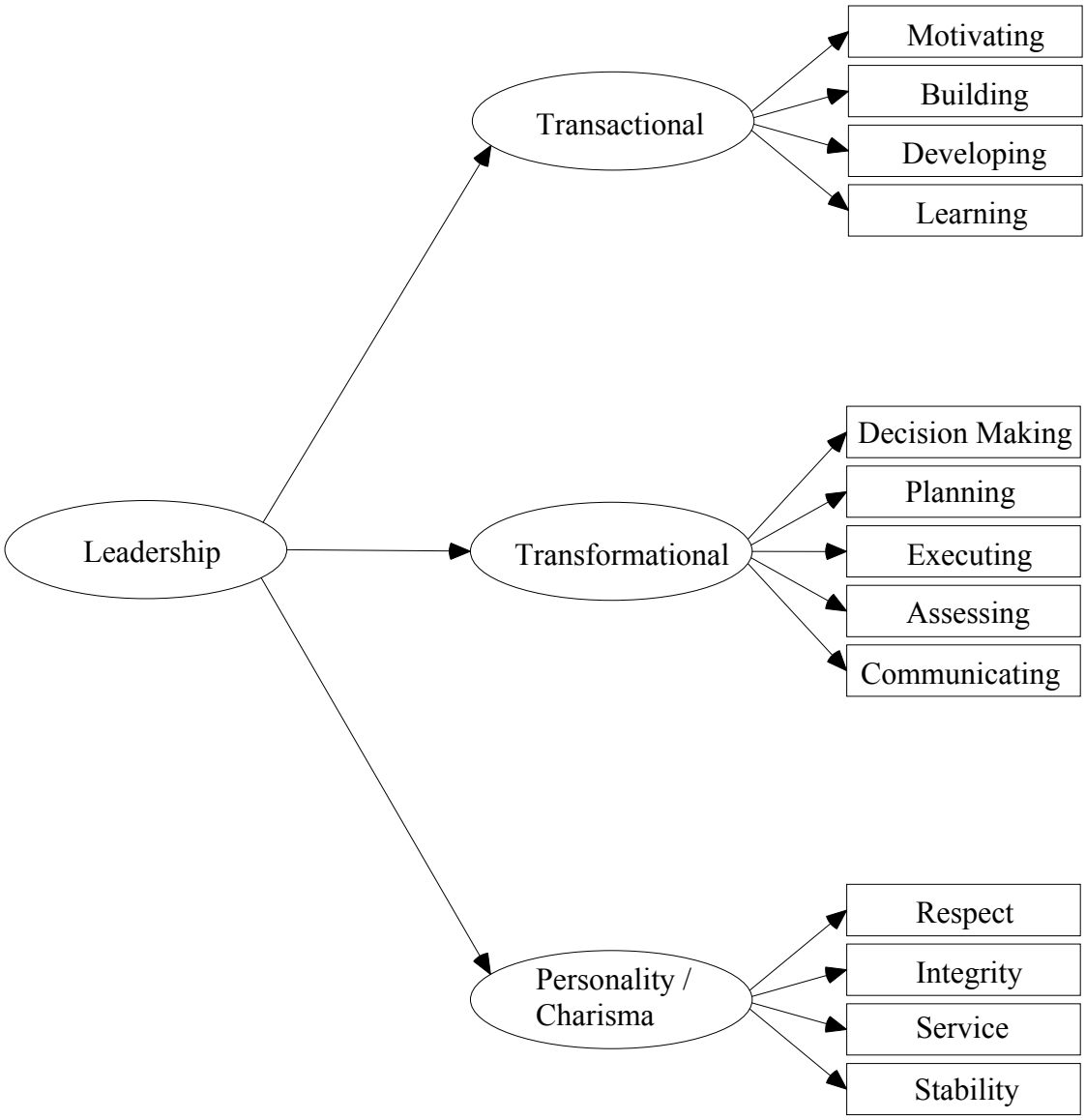
Have Not Observed
Extremely Poor Description
Very Poor Description
Slightly Poor Description
Slightly Good Description
Very Good Description
Extremely Good Description

Executing		
	39. Completes assigned tasks to standard.	○○○|○○○|○
	40. Meets timelines developed to guide work of the staff group.	○○○|○○○|○
	41. Does whatever is necessary (within ethical limits) to complete the mission.	○○○|○○○|○
	42. Monitors execution of plans to identify problems.	○○○|○○○|○
	43. Refines plans to exploit unforeseen opportunities.	○○○|○○○|○
Assessing		
	44. Assesses the staff group's strengths accurately.	○○○|○○○|○
	45. Assesses the staff group's weaknesses accurately.	○○○|○○○|○
	46. Constructively participates in after-action reviews.	○○○|○○○|○
	47. Takes time to find out what other team members are doing.	○○○|○○○|○
Respect		
	48. Actively supports equal opportunity for all persons.	○○○|○○○|○
	49. Creates a climate of fairness in the staff group.	○○○|○○○|○
	50. Excludes some from team activities.	○○○|○○○|○
	51. Treats others with respect.	○○○|○○○|○
Selfless-Service		
	52. Claims credit for others' work.	○○○|○○○|○
	53. Considers the needs of others before self.	○○○|○○○|○
	54. Places the welfare of the staff group before own personal gain.	○○○|○○○|○
	55. Takes privileges not allowed others.	○○○|○○○|○
Integrity		
	56. Behaves with questionable ethics.	○○○|○○○|○
	57. Demonstrates moral courage (does what is right).	○○○|○○○|○
	58. Is not sensitive to the ethical impacts of decisions.	○○○|○○○|○
	59. Is trustworthy.	○○○|○○○|○
	60. Sets the proper ethical example for others.	○○○|○○○|○
Stability		
	61. Displays extreme anger.	○○○|○○○|○
	62. Exhibits wide mood swings.	○○○|○○○|○
	63. Maintains calm disposition under stress.	○○○|○○○|○
	64. Possesses an even temperment.	○○○|○○○|○
	65. Behaves unpredictably.	○○○|○○○|○
Other		
	66. Demonstrates appropriate level of knowledge about the Army.	○○○|○○○|○
	67. Demonstrates appropriate level of branch-specific knowledge and skills.	○○○|○○○|○
	68. Is a clear thinker.	○○○|○○○|○
	69. Maintains effective interpersonal relations with others.	○○○|○○○|○
	70. Sets the example for physical fitness.	○○○|○○○|○
	71. Is a good leader.	○○○|○○○|○
	72. Is someone I would follow into combat.	○○○|○○○|○

For additional information concerning the Leader AZIMUTH Check, contact your CAS3 staff group instructor or Dr. Stanley M. Halpin, Chief, Fort Leavenworth Research Unit, Army Research Institute for the Behavioral and Social Sciences, Halpins@leav-emh1.army.mil.

DesignExpert™ by NCS Printed in U.S.A. Mark Reflex® EW-220204-1 :1098 HR04

Appendix B. ARI path network diagram of the LAC



Note: Adapted from Steele, 2007.

Appendix C. Modified Leadership Azimuth Check items by scale

Communicating

1 Provides clear direction.
2 Explains own ideas so that they are easily understood.
3 Keeps others well informed.
4 Listens well.
5 Tells it like it is.
6 Communicates poorly.

Decision-Making

7 Delays decisions unnecessarily.
8 Generates innovative solutions to unique problems.
9 Ignores information that conflict with own initial assumptions.
10 Makes sound decisions.
11 Willing to revisit a decision with new information calls for it.
12 Effectively uses SOS problem solving techniques.

Motivating

13 Provides good explanation for rationale with directing tasks.
14 Contributes to a supportive environment.
15 Inspires people to do their best.
16 Quick to acknowledge good performance of others.
17 Sets clear performance expectations.

Developing

18 Encourages professional growth.
19 Is an effective teacher.
20 Provides honest feedback to others on their strengths and weaknesses.
21 Sets the example for others by doing his or her best.
22 Seldom shares responsibility with others.
23 Actively participates in the activities of the group.

Building

24 Encourages cooperation among group members.
25 Solicits group input in decision-making situations.
26 Focuses the group on mission accomplishment.
27 Treats others as valuable team members.

Learning

28 Becomes defensive when given critical feedback.
29 Encourages open discussion to improve the group.
30 Helps the group adapt to changing circumstances.
31 Seems to be realistic about own personal limitations.
32 Is willing to accept new challenges.

Planning & Organizing

33 Anticipates how different plans will look when executed.
34 Develops effective plans to achieve group goals.
35 Leaves key events to chance.
36 Sets clear priorities.
37 Is unwilling to modify original plan when circumstances change.
38 Manages time effectively.

Executing

39 Completes assigned tasks to standard.
40 Meets timelines developed to guide work of the group.
41 Does whatever is necessary (within ethical limits) to complete the mission.
42 Monitors execution of plans to identify problems.
43 Refines plans to exploit unforeseen opportunities.

Assessing

44 Assesses the group's strengths accurately.
45 Assesses the group's weaknesses accurately.
46 Constructively participates in after-action reviews.
47 Takes time to find out what other team members are doing.

Respect

48 Actively supports equal opportunity for all persons.
49 Creates a climate of fairness in the group.
50 Excludes some from team activities.
51 Treats others with respect.

Service

52 Claim's credit for other's work.
53 Considers the needs of others before self.
54 Places the welfare of the group before own personal gain.
55 Takes privileges not allowed others.

Integrity

56 Behaves with questionable ethics.
57 Demonstrates moral courage (does what is right).
58 Is not sensitive to the ethical impacts of decisions.
59 Is trustworthy.
60 Sets the proper ethical example for others.

Stability

61 Displays extreme anger.

62 Exhibits wide mood swings.

63 Maintains calm disposition under stress.

64 Possesses an even temperament.

65 Behaves unpredictably.

Other

66 Demonstrates appropriate level of knowledge about the Air Force.[1]

67 Demonstrates appopriate level of AF specialty-specific knowledge and skills.[2]

68 Is a clear thinker.

69 Maintains effective interpersonal relations with others.

70 Sets the example for physical fitness.

71 Is a good leader.

72 Is a good Air Force officer.[3]

Notes: [1] The original question uses "Army" rather than "Air Force." [2] The original question uses "branch-specific" rather than "AF specialty-specific." [3] Changed from "Is someone I would follow into combat" to be more inclusive of the various AFSCs attending SOS.

Appendix D. Modifed LAC used in SOS survey administration

AU SCN 08-001

This questionnaire consists of a number of statements that are designed to understand different thoughts and attitudes that people have and behaviors they enact.

DIRECTIONS: Use the scale below to indicate to what extent you agree with the following statements.

Strongly Disagree 1	Disagree 2	Neutral 3	Agree 4	Strongly Agree 5	
1. The concepts and skills taught in SOS are important to me.					
2. SOS provides a good environment in which to learn leadership skills.					
3. SOS attendance is important to my personal development.					
4. The lessons presented in SOS are effective in engaging me.					
5. I will be able to apply the lessons learned in SOS in my job.					

DIRECTIONS: Use the scale below to indicate to what extent you believe each statement describes yourself and the three people you are rating. In your assessment, please compare the person you are rating with others you have known well.

Not at All 1	To a Small Extent 2	Moderately Well 3	Quite a Bit 4	To a Great Extent 5	Self	08AF6006S	08AF6009S	08AF6005S
6. Does not address unethical behavior, even though he/she has noticed it.								
7. I cannot see this person undertaking a promising but risky new business venture.								
8. Encourages open discussion to improve the group.								
9. Would confront the inappropriate behavior of his/her superiors.								
10. Is a good Air Force officer.								
11. Is an effective teacher.								
12. People consider this person to be a courageous person.								
13. Does whatever is necessary (within ethical limits) to complete the mission.								
14. Is unwilling to modify original plan when circumstances change.								
15. Actively participates in the activities of the group.								
16. Delays decisions unnecessarily.								
17. Claim's credit for other's work.								
18. Treats others as valuable team members.								
19. Would not tolerate behavior by someone, even a friend, that hurts others and would confront the person about the behavior.								
20. People do not think of this person as being courageous.								
21. Believes that risking your own safety to save somebody else is better than not acting and risking somebody else getting hurt.								
22. When this person sees somebody act unethically, he/she confronts the person with what they have done.								
23. Quick to acknowledge good performance of others.								
24. Provides clear direction.								
25. Generates innovative solutions to unique problems.								
26. Possesses an even temperament.								
27. I do not consider this person to be courageous.								
28. Displays extreme anger.								

Not at All (1)	To a Small Extent (2)	Moderately Well (3)	Quite a Bit (4)	To a Great Extent (5)	Self	08AF6006S	08AF6009S	08AF6005S
29. Actively supports equal opportunity for all persons.								
30. Speaks up when workplace happenings conflict with his/her sense of what is appropriate.								
31. Maintains effective interpersonal relations with others.								
32. In the workplace, this person pursues promising new ideas, even though they may not work out.								
33. Is willing to accept new challenges.								
34. Encourages professional growth.								
35. Takes time to find out what other team members are doing.								
36. Would do what he/she could to save a stranger's life, even if he/she were to risk an injury to him/herself.								
37. Encourages cooperation among group members.								
38. Creates a climate of fairness in the group.								
39. Helps the group adapt to changing circumstances.								
40. Would stand up to a supervisor or higher level manager if they asked him/her to do something he/she was not sure was ethical.								
41. Seldom shares responsibility with others.								
42. Maintains calm disposition under stress.								
43. Is likely to remain silent about a peer's ethics violation.								
44. Becomes defensive when given critical feedback.								
45. Sets clear performance expectations.								
46. Treats others with respect.								
47. Refines plans to exploit unforeseen opportunities.								
48. Would risk getting hurt to save somebody else from getting hurt.								
49. Assesses the group's strengths accurately.								
50. Meets timelines developed to guide work of the group.								
51. Is trustworthy.								
52. Is more likely than others to make personal physical (in the form of risking safety) sacrifices to help others.								
53. Sets the proper ethical example for others.								
54. Demonstrates appropriate level of knowledge about the Air Force.								
55. Exhibits wide mood swings.								
56. Excludes some from team activities.								
57. Is not sensitive to the ethical impacts of decisions.								
58. Manages time effectively.								
59. Inspires people to do their best.								
60. Listens well.								
61. Leaves key events to chance.								
62. Sets the example for others by doing his or her best.								
63. Makes sound decisions.								
64. Behaves with questionable ethics.								
65. Focuses the group on mission accomplishment.								
66. Is a good leader.								

Not at All 1	To a Small Extent 2	Moderately Well 3	Quite a Bit 4	To a Great Extent 5	Self	08AF6006S	08AF6009S	08AF6005S
67. Effectively uses SOS problem solving techniques.								
68. Solicits group input in decision-making situations.								
69. Willing to revisit a decision with new information calls for it.								
70. Explains own ideas so that they are easily understood.								
71. Develops effective plans to achieve group goals.								
72. Assesses the group's weaknesses accurately.								
73. Seems to be realistic about own personal limitations.								
74. Has a low tolerance for inappropriate behavior and would act to stop it.								
75. If this person were in a position where they noticed somebody act unethically, he/she would address the problem regardless of repercussions for him/herself.								
76. Demonstrates moral courage (does what is right).								
77. Places the welfare of the group before own personal gain.								
78. Takes privileges not allowed others.								
79. Provides honest feedback to others on their strengths and weaknesses.								
80. Provides good explanation for rationale with directing tasks.								
81. If somebody needed to be rescued, this person would act immediately even though he/she may feel some fear.								
82. Believes that acting on a good idea for an organizational innovation while risking your own reputation is better than not acting on the idea and risking the performance of the organization.								
83. Monitors execution of plans to identify problems.								
84. Is a clear thinker.								
85. Sets clear priorities.								
86. Keeps others well informed.								
87. Would take a stance to help physically protect others if they were in danger.								
88. Constructively participates in after-action reviews.								
89. Ignores information that conflict with own initial assumptions.								
90. Contributes to a supportive environment.								
91. Demonstrates appropriate level of AF specialty-specific knowledge and skills.								
92. Behaves unpredictably.								
93. Communicates poorly.								
94. Anticipates how different plans will look when executed.								
95. Tells it like it is.								
96. Sets the example for physical fitness.								
97. Completes assigned tasks to standard.								
98. Considers the needs of others before self.								

THANK YOU FOR YOUR PARTICIPATION

Notes: As administered, the instrument included 72 modified LAC items and 21 items from an instrument investigating courage development. The questions of each instrument are interspersed.

Appendix E. SOS leadership & management instruction modules

When accomplished	Comparison Group	Treatment Group
Prior to pretest	S2130-Evaluation	S2130-Evaluation
	S2120-Teambuilding	S2120-Teambuilding
	S2330- Individual Decision Making & Goal Setting	S2330- Individual Decision Making & Goal Setting
	S2230-APTEC Seminar	S2230-APTEC Seminar
	S2340-Team decision making & conflict management	S2340-Team decision making & conflict management
	S2350-Team structure & culture	S2350-Team structure & culture
	S3210- Followership	S3210- Followership
	S2510-Mentoring & developing Airmen	S2510-Mentoring & developing Airmen
	S2320-Situational Leadership	S2320-Situational Leadership
	S2325-Situational leadership case studies	
	S2515-Reflections on developmental counseling	
Between administrations	S2900-Leadership guest speaker	S2325-Situational leadership case studies
		S2515-Reflections on developmental counseling
		CLX
		S2900-Leadership guest speaker

Note: The leadership guest speaker, the commander of Air University, was the same for both groups.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From – To)*
27-03-2008	**Master's Thesis**	Sep 2006 - Mar 2008

4. TITLE AND SUBTITLE
Evaluating Experiential Leader Development: A Programmatic Evaluation and Comparison of the Effectiveness of US Air Force Squadron Officer School Curricula

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Holland, Jeffrey G., Captain, USAF

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S)

Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/EN)
2950 Hobson Street, Building 642
WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT/GLM/ENV/08-M01

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

SOC/DES (AETC)
Attn: Dr. Arden Gale
Bldg 1403, Rm 2257
125 Chennault Circle DSN: 953-9436
Maxwell AFB AL 36112-6417 e-mail: arden.gale@maxwell.af.mil

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Leader development programs often employ experiential learning exercises. The impact of such exercises is not clear. This research investigated experiential leader development using a quasi-experimental design to analyze the differences in two consecutive US Air Force Squadron Officer School (SOS) in-residence classes. The curriculum was altered between classes by the addition of the Combat Leadership Exercise (CLX), an experiential war-gaming activity. Experiential programs regularly use mean differences between pretest and posttest measurements to represent program impact. However, research shows that participants may change the way they evaluate themselves between test administrations due to their experiences in the programs, a phenomenon known as response shift. Response shift renders results of mean differences evaluation invalid. The common means differences showed SOS had weak impact on leader development and showed no difference between the treatment class (CLX) and the comparison class (no CLX). However, structural equation modeling identified the presence of response shift within each SOS class, indicating that students had reconceptualized or recalibrated certain aspects of leadership measured before and after SOS. The implications of response shift and its measurement are discussed. An argument for changing the leader development evaluation paradigm to legitimize response shift as a program outcome is presented.

15. SUBJECT TERMS
Leadership, Leadership Development, Leadership Training, Management Training, Transfer of Training, Human Resources, Experiential Learning

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Kent C. Halverson, Lt Col, USAF (ENS)
U	**U**	**U**	**UU**	94	**19b. TELEPHONE NUMBER** *(Include area code)* (937) 255-3636, ext 4709; e-mail: Kent.Halverson@afit.edu

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39-18